PB-250 638

REALISTIC MODELS FOR MORTALITY RATES AND THEIR ESTIMATION

V. K. Murthy

California University

Prepared for:

Health Effects Research Laboratory

February 1976

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA-600/1-76-015	2.
3. RECIPIENT'S ACCESSION NO.	
4. TITLE AND SUBTITLE: REALISTIC MODELS FOR MORTALITY RATES AND THEIR ESTIMATION	5. REPORT DATE: February 1976
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S): V. K. Murthy	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of California at Los Angeles and University of Southern California, Los Angeles, California	10. PROGRAM ELEMENT NO.: 1AA601
	11. CONTRACT/GRANT NO.: 800230
12. SPONSORING AGENCY NAME AND ADDRESS: Health Effects Research Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Research Triangle Park, N.C. 27711	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE: EPA-ORD
15. SUPPLEMENTARY NOTES	

16. ABSTRACT

The objective of a medical follow-up study is generally to determine the effectiveness of each of several treatments by analyzing the responses of the patients. Frequently the response data coming out of these investigations is time to death of patients who are not otherwise lost to the follow-up of our investigation. The statistical nature of this data are characterized in this report.

By definition the, "Force of mortality or mortality rate function", is the rate associated with the probability of the patients' death in a specified short interval of time, given that the patient has survived to this instant in time.

Mathematical models are presented which, as special cases, represent constant, increasing and decreasing mortality rates, along with combinations of these properties. Usually, these mortality rate curves are "U" shape. The first part of the curve corresponds to infantile mortality, the second part corresponds to useful iife, and finally, the last part corresponds to decay, aging, etc., culminating in death. Their corresponding probability distributions and survivorship functions are obtained in closed form. Methods of estimating the parameters are developed and procedures dealing with computational details and statistical properties are discussed.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Mortality Mathematical Models		12A 05B 06G

PRICES SUBJECT TO CHANGE

18. DISTRIBUTION STATEMENT: RELEASE TO PUBLIC	19. SECURITY CLASS *(This Report)*: UNCLASSIFIED	21. NO. OF PAGES: 106
	20. SECURITY CLASS *(This page)*: UNCLASSIFIED	22. PRICE: 5.50-2.25

EPA Form 2220-1 (9-73)

015072

PB 250 638

EPA-600/1-76-015
February 1976

Environmental Health Effects Research Series

MORTALITY RATES [illegible]

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U. S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

RESEARCH REPORTING SERIES

Research reports of the Office of Research and Development, U.S. Environmental Protection Agency, have been grouped into five series. These five broad categories were established to facilitate further development and application of environmental technology. Elimination of traditional grouping was consciously planned to foster technology transfer and a maximum interface in related fields. The five series are:

1. Environmental Health Effects Research
2. Environmental Protection Technology
3. Ecological Research
4. Environmental Monitoring
5. Socioeconomic Environmental Studies

This report has been assigned to the ENVIRONMENTAL HEALTH EFFECTS RESEARCH series. This series describes projects and studies relating to the tolerances of man for unhealthful substances or conditions. This work is generally assessed from a medical viewpoint, including physiological or psychological studies. In addition to toxicology and other medical specialities, study areas include biomedical instrumentation and health research techniques utilizing animals—but always with intended application to human health measures.

This document is available to the public through the National Technical Information Service, Springfield, Virginia 22161.

EPA-600/1-76-015
February 1976

REALISTIC MODELS FOR MORTALITY RATES AND THEIR ESTIMATION

By

V. K. Murthy
University of California at Los Angeles
and
University of Southern California
Los Angeles, California

Grant No. 800230

Project Officer

Dr. Wilson B. Riggan
Population Studies Division
Health Effects Research Laboratory
Research Triangle Park, N.C. 27711

U.S. ENVIRONMENTAL PROTECTION AGENCY
OFFICE OF RESEARCH AND DEVELOPMENT
HEALTH EFFECTS RESEARCH LABORATORY
RESEARCH TRIANGLE PARK, N.C. 27711

DISCLAIMER

This report has been reviewed by the Health Effects Research Laboratory, U.S. Environmental Protection Agency, and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the U.S. Environmental Protection Agency, nor does mention of trade names or commercial products constitute endorsement or recommendation for use.

ABSTRACT

The object of a "medical follow-up study" is generally to determine the effectiveness of each of several types of treatments or regimen by analyzing the responses of the patients. Depending on the nature of the disease and the nature of the drugs or combination of drugs or other forms of treatment such as surgery chemotherapy etc. a "Protocol" is developed with the sole object of rendering the study feasible and valid. Thus the "Protocol" concept is much broader in scope than the concept of plain so-called clinical trials which often give the investigator the misleading belief that he has an experimental design (in its usual sense) for his study as in the case of agricultural and animal experimentation.

The most reliable reponse data coming out of these investigations thus far is time to death of patients who are not otherwise lost to the follow-up or our investigation. The statistical nature of this data among other things, will be characterized in the body of this report. A salient parameter of interest is the force of mortality experienced by patients under the specific protocol.

By definition the force of mortality is the rate associated with the probability of the patient's death at a specified neighborhood of time given that the patient has survived that instant of time. This quantity expressed a a function of time is also called the "mortality rate function." Given the mortality rate function, the survival function can be determined.

The estimation of the mortality rate function therefore assumes an important role in the estimation as well as prediction, of the survival function. Nonparametric methods for estimating hazard rates are considered in Murthy (1965) and Grenander (1956). However, the most that the aforementioned rates can do is to throw light on the shape of the underlying hazard rate function, since they cannot be used for decision making

and prediction purposes, as the estimation of a continuous infinity of parameters is implicit in all those procedures. Thus, parametric procedures based on realistic models for the mortality rate are the only suitable approaches for efficient estimation and prediction. It is generally accepted, based on both deductive reasoning and extensive empirical evidence, that the ideal mortality rate curve consists of an early decreasing phase due to burn-in and de-bugging (infantile mortality) followed up by a phase of useful life, where it remains approximately constant before climbing to failure due to wear out, aging, or fatigue damage. However, no models for the mortality rate have been advanced so far that depend on as few parameters as possible and which also contain all the three phases mentioned earlier and which at the same time, give rise to the survival function in a simple closed form.

Therefore families of distributions which contain those corresponding to the increasing mortality rate (IMR), those corresponding to the decreasing mortality rate (DMR), and finally, those corresponding to the constant mortality rate (exponential distributions) and combinations thereof giving rise to the classical bath-tub shape mortality rate curves, are characterized in a closed form in this report. Methods of estimating the parameters are developed and procedures dealing with computational details and statistical properties are discussed.

A graphics program is developed which gives rise to the curve that is most apt, in a given situation, by playing with the parameters in the present models. The interactive part of the program, as well as the problem of dealing with competing risks in a realistic manner are reserved for the continuation part of the study.

TABLE OF CONTENTS

Chapter		Page
1	Introduction	1
2	Annotated Bibliography on Statistical Analyses of Follow-Up Studies for Patients	7
3	Realistic Models for Mortality Rates and Their Estimation	38
	Introduction	38
	Estimation of the Hazard Rate Function	38
	Theory	39
	Realistic Models of the Mortality Rate	42
	Application to Medical Follow-Up Studies	45
	Life Table Model	49
	Follow-Up Model	49
	Test Results	59
4	Applications of the Models and an Example	72
	Real Life Examples to Which the Model Can be Applied	72
	Statistical Models with the Bath-Tub Property	73
	Graphics Program	74
	Maximum Likelihood Procedure for Estimation	75
	A Numerical Example	81
5	Computer Program for Graphics	84
6	Concluding Remarks	94
	Bibliography	95

CHAPTER 1
INTRODUCTION

The investigation leading to the study in this final report interestingly, has its roots in a typical problem which is widely encountered in medical or epidemiological follow-up studies. The particular follow-up study, brought to the attention of this author via Dr. Emanuel Landau then at the Bureau of Radiological Health, and Dr. W. J. Dixon of the Health Sciences Computing Facility at U. C. L. A., is widely known as "The Cooperative Thyro-Toxicosis Study." The data base came from twenty-four hospitals in the United States and one from the United Kingdom. Several people notable Mrs. Tompkins, presently of the EPA at Durham, North Carolina and Professor Lewis of the California Institute of Technology, Pasadena, California have worked with this data base and came up with conflicting conclusions:

Dr. Lewis maintained that the incidence of the acute form of leukemia in patients treated with radioactive iodine is statistically significantly larger than the incidence among those treated with tyrodactomy and/or medicines.

On the other hand, Mrs. Tompkins maintained, basing her analyses on the same data base, that there was indeed no significant difference in the incidence of the acute form of leukemia among the two groups. The author agrees with the stand taken by Mrs. Tompkins because of the followingtwo flaws in the arguments advanced by Dr. Lewis:

The main flaw is the age grouping that Dr. Lewis has resorted to (indicating possible deliberateness to make the data vouch-safe to the stand taken by him) lies in its arbitrariness rather than taking it as a concomittant or co-influencing variable.

The other flaw is inherent to the data base by way of the relatively infinitesimally small incidences of the acute form of leukemia in relation to the large sample of the patients.

Also, the empirical age distributions of the patients in the two groups, namely those treated with radioactive Iodine and those treated otherwise, turn out to be almost the mirror images of one another, the one corresponding to those subject to surgery being skewed to the right while the other, being its mirror image is skewed to the left.

More explicitly and in greater detail we recall the following findings:

The February 1972 issue of the Journal of the National Cancer Institute contained an article entitled "Irradiation in the Epidemiology of Leukemia Among Adults," by Gibson et al. This paper was also presented by them at the American Public Health Association annual meeting in October 1971. The authors stated in the summary that the relation between leukemia and diagnostic medical irradiation was compared. The comparison was based on 1,414 cases of adult leukemia in patients who were diagnosed as having leukemia or who died with leukemia between 1959 and 1962 in twenty-six upstate New York counties and counties in and around Baltimore, Maryland, and Minneapolis, St. Paul, Minnesota. These 1,414 cases were compared with a so-called control group of 1,370 persons randomly selected from households in the same geographic area as the cases in the sample population. The association of irradiation with leukemia was examined by histologic type sex, and age of the subject. The two histoligic types of leukemia considered were lymphatic and myeloid-monocytic and within each type the acute and chronic types were distinguished. In the rest of the cases the type of leukemia was unspecified.

From a statistical analysis of the data, the authors concluded that:

(a) irradiation had an enhanced risk for both acute and chronic myeloid leukemia in males but not in females;

(b) the effect or association was more pronounced for irradiation to the trunk than to all other sites;

(c) neither irradiated males nor females had an elevated risk

for any other histologic type of leukemia; and

(d) only a small proportion of the irradiated population developed leukemia, and irradiation accounted for only a small proportion of leukemia cases.

Many criticisms can be made about their report and procedures. There are not sufficient data published to examine their adjustments for age differences in the exposed and control groups. Any modification could affect the analysis significantly. The computations for estimated relative risk were not given. Examination of these computations would be desirable since the observed correlation of leukemia with number of films throughout their tables occur, whether the ratios are significantly large or not. This may indicate an even stronger relation between leukemia and exposure than these authors indicate.

Some of the fundamental criticisms about most of the analyses made by different investigators of the thyro-toxicosis data base were given in a paper by Murthy, et al, that appeared in the Journal of the National Cancer Society. Their criticisms are:

(1) The major defect of the paper is the large difference in age distribution between cases and controls. Only 37 percent of the controls were over sixty years of age compared with 62 percent of the cases. An age standardization may not be an adequate compensation for such a large difference, in view of the implied cohort differences. Moreover, the chi-square test is a crude assessment of the proportional differences of cases and controls with a given attribute, and the Cochran modification cannot be relied on to control variations in the age differences of the two groups.

(2) So much of the paper deals with the method of removing possible sources of bias from the data that it is difficult if not impossible to characterize the ultimate samples which form the basis of the author's analysis.

(3) Table 1 on page 303 shows age distribution (in 10-year intervals) of 1, 370 controls and 1960 census data for the appropriate geographic area which includes approximately 13 million persons. These data

are not consistent with the statement on page 302 that the controls form approximately a "one in 3,000 probability sample" of the population. Also, in our opinion a "probability sample" is not relevant unless the authors of the paper explain their meaning of the term.

4) In our opinion, the paper would be strengthened by a statistical appendix, explaining the age adjustment, the weighted percent exposed, the test of significance, and the estimation of relative risk; if variances are important, they too should be included.

5) The suggested appendix should also

   a) describe the respondents for cases and controls, to determine whether husbands speaking for their wives and vice versa, are equally reliable, and

   b) show the effectiveness of the "auxiliary reporting system," e.g., did this affect the proportions of cancers notified during life and after death? If, as the reviewers suggest, a high proportion of leukemias was first reported by death certificates, were the radiographic histories equally reliable for leukemias and other cancers?

6) Conclusions comparing trunk with all <u>other</u> sites, as stated by the authors, cannot be examined, since they merely report "all sites" and "trunk only."

7) The conclusion, that X-rays of the chest and abdomen were responsible for a small proportion of the myeloid leukemias, did not consider all the pertinent factors. For instance, the authors did not record intervals between exposure and onset or make allowance

for the possibility that the preleukemic state, which increases susceptibility to infection, might increase the chance of being X-rayed during the 5 - 6 years before the onset of symptoms.

The interval between exposure and onset is important. Stewart, in his article "Tissue Aging as a Factor in Juvenile Cancers," [Proc. R. Soc. Med., 65-245-246, 1972] has shown that when myeloid leukemia is caused by obstetric radiography, the interval between birth and onset of symptoms is usually 5 - 10 years. The preleukemic state is also important. The probability of developing a nonfatal respiratory infection while incubating leukemia is increased, as described in Kneale's article "Excess Sensitivity or Pre-Leukemics to Pneumonia: A Model Situation for Studying the Interaction of Infectious Disease with Cancer," [Br. Prev. Soc. Med., 25:152-159, 1971], as well as the probability of developing a fatal infection. The latter is brought up in Kneale's 1971 article as well in a study done by A. Stewart, "Epidemiology of Acute (and Chronic) Leukemias," appearing in Clinics in Haematology, edited by S. Roath [London: W.B. Saunders, 1972].

It was suggested by Stewart, et al, in an article entitled "Adult Leukaemias and Diagnostic X-rays," [Br. Med. J., 5309:882-890, 1962] that X-rays of the chest and abdomen might be responsible for a few myeloid leukemias in adults. However, their data showed that only for five years before the onset of symptoms were the patients with myeloid leukemia more at risk of being X-rayed than their controls. Therefore, it is reasonable to conclude that the extra X-ray examinations in the adult surveys were a consequence of

the heightened sensitivity of the recipients and not a cause of the disease.

Since basic issues such as the selection of an appropriate model, the procedure used to adjust for the age of the subject, and finally, taking into consideration the environmental factors and the subject's individual medical status are involved in all these studies, the author undertook a systematic investigation that will probably extend over a period of time, in order to tackle these basic issues.

Towards this end, during the first phase of this investigation and culminating in this report, the author made an extensive study of the state of the art of procedures for analysis, under the title "Annotated Bibliography. This has been summarized in Chapter Two of this final report. Chapters Three and Four deal with extremely versatile models and an example of the relative performance of the models dealing with "Survival Data," dealing with a hypothetical cohort of 1,000 people born in the United States in the year 1900, is given.

In Chapter 5 a graphics method which can be made interactive is described and this procedure will enable the investigator to choose the appropriate model, based on his empirical experience.

In Chapter Six, the author points out how these methods can be applied to assess the marginal detrimental effect of any designated pollutant on the mortality rate, in the presence of other competing risks such as the concentrations of an entire spectrum of harmful pollutants.

Annotated Bibliography on Statistical Analyses of Follow-Up Studies for Patients

In this chapter, we review comprehensively the state of the art situation in modeling for the survival curve and the methods used for comparing different survival curves. The review follows the alphabetical listing of the authors. This chapter, therefore, gives an insight into where we stand in our ability to deal with the survival experiences of patients in the (competing risk) environment subjected to different regimens or treatments. Consequently, this chapter serves as a stimulant for additional research, meeting the real world situation and revealing challenges as to how the investigator should tackle the various violations of the underlying design or profile.

In his paper "The Comparison of Survival Curves," Armitage [1959] considered the problem of two survival curves. The survival time distributions, $F_A(t) = 1 - e^{-\lambda t}$ and $F_B(t) = 1 - e^{-\lambda' t}$, are compared by the statistic

$$\chi = \frac{(\lambda' - \lambda)}{\lambda} \sqrt{n\psi(\lambda T)} \ .$$

Several techniques are compared by considering the ratio of their respective $\psi(\lambda T)$'s.

Parametric maximum likelihood estimators of λ are

$$\frac{1}{\hat{\lambda}} = \frac{\Sigma T_i + \Sigma t_j}{d} \ ,$$

where t_j, $1 \leq j \leq d$, is the time at death from entry and T_i, $1 \leq i \leq n - d$, are the survivors' times since entry out of a total of n patients entering treatment.

The Sign Method assumes that patients enter in and one is given treatment A and the other is given treatment B. This test requires that $(1 - F_A(t)) = (1 - F_B(t))^k$ for some constant k. Any of the sequential procedures for binomial sequences can be applied.

Direct comparison of proportions of survivors at time τ after treatment is shown to be less efficient than the sign method.

Actuarial methods divide the interval $(0,\tau)$ into equal subintervals and estimate the probability of survival through each subinterval directly. The product-limit (PL) estimate is

$$\hat{P}(\tau) = \prod_{r} \frac{n - r}{n - r + 1} ,$$

where r is the number of patients for which $tr \leq \tau$. This is a non-parametric maximum likelihood estimator of $F(t)$.

Computations of relative efficiencies are given and an illustrative example is worked out.

A "Study of Life-Testing Models and Statistical Analysis for the Logistic Distribution," has been made by Bain, et al, [1973] in which they have reviewed bath-tub hazard rates in Section I of their paper; the model I

of Murthy and Swartz [1971] is shown to correspond to the distribution of the minimum of a Pareto and a Weibull variable. It is shown that the minimum of Weibull variables gives a hazard rate

$$h(t) = \alpha_1\beta_1 t^{\beta_1-1} + \alpha_2\beta_2 t^{\beta_2-1}$$

which is U-shaped in $\beta_1 < 1$ and $\beta_2 > 1$.

Estimation of the parameters in a polynomial model are discussed in Section II. For a linear hazard, as proposed by Kodlin [1967], it is shown that the maximum likelihood equation $\partial L/\partial a = 0$ and $\partial L/\partial b = 0$, where $L = L(t_1,\ldots,t_n ; a,b)$ in the likelihood function may have multiple solutions and that the maximum likelihood may occur on the boundary. The Cramér-Rao lower bounds are computed for sample sizes of 10, 20, 40 and 80. Estimates using a least squares fit to the empirical distribution function and a fit of $\ln (1 - F(t_i))$ to $\ln (1 - (i/n+1))$ are developed. A test of $b = 0$ vs. $b \neq 0$ is proposed for testing whether the distribution is exponential or has a linear increasing hazard rate.

Section III presents the results of Monte Carlo techniques to obtain the distributions of maximum likelihood estimates of the logistic distribution.

In a paper dealing with the "Failure of a Model to Predict Cancer Survival," Berg and Robbins [1967] consider that the mortality due to cancer is assumed to be constant and equal to a linear sum of tumor size, tumor

grade, node metastases and node sinus histiocytosis. It is shown that for a time beyond ten years after mastectomy, mortality does not agree with the assumption of constant mortality.

Following this, Berkson and Gage [1952] take into consideration a "Survival Curve for Cancer Patients Following Treatment," using a function $l_t = cl_0 + (1-c)l_0 e^{-\beta t}$, of years after operation, t, depending upon two parameters. These parameters are c the percentage cured and β the death rate from the cancer. It is fitted by least squares to data on survival of cancer of the stomach. The function is based on the assumption that a certain fraction of the cancerous population is dying at the normal rate, while the rest encounter an instantaneous rate of risk to the cancer that is constant. The function fits the data very closely. It is recommended that c,β, with the ratio of expectations of life of the cancerous to the normal population be used as a summary of the mortality of the cancer, rather than the five-year survival rate. Adjustment for age-at-entry-to-treatment is not made.

Boag [1949] uses maximum likelihood estimates to obtain the proportion of patients cured by cancer therapy. Histograms of survival times for all patients who died from cancer, after treatment, show a distribution that is skewed with a marked tail extending towards long survival times. This is true for data collected from various hospitals, for different cancers, various sites of the disease and different methods of treatment. The gamma, Weibull, lognormal and exponential distributions were considered, but the

lognormal gives the best fit, as measured by chi-square goodness-of-fit criterion. For a large group of people treated for cancer, the survival time is described by a distribution that is a mixture of c percent who are permanently cured and $(1-c)$ percent whose survival time follows a lognormal distribution with parameters μ and σ.

Three estimation cases are discussed: estimation of μ, σ and c; estimation of μ and c with σ unknown; and estimation of c with μ and σ known. Methods of numerically solving the maximum likelihood equations are given and a numerical example, based on data from 121 cases of carcinoma of the breast, is given.

The amount of information of the estimator, c, which is the reciprocal of its variance, is determined as a function of the number of new patients in a study and T is the total duration of the study. This is made under the assumption that death occurs only as a result of cancer and not from other causes and that no patients with a recurrence of the cancer are alive at the end of the test period. The three situations listed above are considered also. The amount of information is obtained when a constant number of patients present themselves for treatment each month.

The ages of the patient, at the time of treatment, is considered by the following model:

$$F(x)^{dx} = \left(\int_0^{\infty} f(y,0)\, g(x+y)\, e^{-\int_0^x g(x'+y)dx'}\, dy \right) dx \ ,$$

where $f(x,y)dx$ is the number of survivors after y years and $f(x,0)$ is the age distribution among patients at the time of treatment, with $F(x)$ the total number of deaths in the interval $(x, x + dx)$ and $g(x + y)$ is the death rate of the normal population. The derivation was made on the assumption that the age of the patient has no effect on the response of the tumor to the treatment.

The discussion contained in this paper brings out many practical considerations involved in follow-up studies such as the time between occurrence of the cancer and its detection, the accurate determination of cause of death, and the assumption of independence of deaths from cancer and deaths from other causes.

In his paper "Simple Mortality Rates," Broadbent [1958] using the mortality rates, $m(t)$, has fitted them by graphical plots to data on the lifetime of milk bottles.

	$m(t)$
Constant	a
Linear	$a + 2b\ t$
Strange	$a^b b\ t^{b-1}$
Gompertz	$a\ e^{bt}$
Makeham	$a\ e^{bt} + c$

Buckland [1964] presents the advantages of using the hazard rate in his "Statistical Assessment of the Life Characteristic. A list of reference papers that consider Weibull, Gamma, Normal, Extreme Value and Log Normal distributions and their estimates are listed. Chapter 2.15 of this paper surveys the literature on the life table.

In his paper, "A Note on the Consistency and Maxima of the Roots of Likelihood Equations," Chanda [1954] presents properties of the maximum likelihood estimators for $\Theta = (\Theta_1,\ldots,\Theta_k)$ unknown parameters which are developed for random samples of size n. The density, $f(x,\Theta)$, is assumed to meet the following regularity conditions:

1. Θ lies in a k-dimensional interval Ω,

$$\frac{\partial \log f}{\partial \log \Theta_r} \quad , \quad \frac{\partial^2 \log f}{\partial \Theta_r \partial \Theta_s} \quad \text{and} \quad \frac{\partial^3 \log f}{\partial \Theta_r \partial \Theta_s \partial \Theta_t}$$

exist for all $r,s,t = 1,2,\ldots,k$ for all $\Theta \in \Omega$ and for almost all x.

2. For almost all x and for all $\Theta \in \Omega$

$$\left| \frac{\partial f}{\partial \Theta_r} \right| < F_r(x) \quad , \quad \left| \frac{\partial^2 f}{\partial \Theta_r \partial \Theta_s} \right| < F_{rs}(x)$$

and

$$\left| \frac{\partial^3 \log f}{\partial \Theta_r \partial \Theta_s \partial \Theta_t} \right| < H_{rst}(x)$$

with

$$\int_{-\infty}^{\infty} H_{rst}(x) f \, dx < M$$

for all $r,s,t = 1,2,\ldots,k$ where M is a positive constant.

3. For all $\theta \in \Omega$ the matrix J,

$$J_{rs}(\theta) = \int_{-\infty}^{\infty} \frac{\partial \log f}{\partial \theta_r} \frac{\partial \log f}{\partial \theta_s} f \, dx$$

is positive-definite and $|J|$ is finite.

By expanding

$$\log \varphi = \sum_{i=1}^{n} \log f_i$$

as

$$L_r(\theta) = L_r(\theta^\circ) - \sum_{s=1}^{k} (\theta_s - \theta_s^\circ) L_{rs}(\theta^\circ) + \frac{1}{2} \sum_{t=1}^{k} (\theta_s - \theta_s^\circ)(\theta_t - \theta_t^\circ) L_{rst}$$

where

$$L_r(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial \log fi}{\partial \theta_r} \right),$$

$$L_{rs}(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial^2 \log fi}{\partial \theta_r \partial \theta_s} \right),$$

and

$$L_{rst} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\partial^3 \log f_i}{\partial \theta_r \partial \theta_s \partial \theta_t} \right)$$

for all $\theta = \theta'(x_i)$, where $\theta'(x)$ is a value depending on x for $r = 1,2,\ldots,k$.

It is shown that $L_{rs}(\theta^\circ) \to J_{rs}(\theta^\circ)$ with probability tending to 1 and that $\theta_r - \theta_r^\circ$ have asymptotically a joint k-variate distribution with zero means and variance-covariance matrix $V = (n\, J_{rs}(\theta^\circ))^{-1}$. Of all possible solutions to the likelihood equations, only one tends, in probability, to the true parameter, θ°.

Chiang [1961] presents a "Stochastic Study of the Life Table and its Applications," where the lifetimes of individuals in a group are assumed to be i.i.d. random variables and it is also assumed that a constant force of mortality for each year applies for each cause of death. The case of censored samples is considered. Maximum likelihood estimators are found for p_x, the probability that an individual who is alive at time, x, will survive the interval $(x, x+1)$, assuming a constant force of mortality at each x. Computation of the life-table is made, using p_x, if all who entered the study die before the study is completed, that is, if life withdrawals occur the life table is approximated by assuming a constant force of mortality beyond the end of the study period.

In the case of simultaneous risks, it is assumed that in each year the

force of mortality is constant and their sum is the total force of mortality. The crude probability of death, in the presence of all risks, the net probability of death due to one cause, and the partial crude probability of death due to a specific cause with all but one of the competing risks eliminated. The maximum likelihood estimators of these probabilities are obtained. The asymptotic variance and covariance of the estimators are given, as is an example of life-table construction using cancer data. The interval over which the force of mortality is constant depends on the investigator; in this case, one year was selected.

It is remarked that the problem of cases lost due to failure to follow up is unsolved and probably has no unique solution. Suggestions for handling the data are given if the number of lost cases is small.

David [1970] has presented a paper on "Chiang's Proportionality Assumption in the Theory of Competing Risks," that an individual's lifetime, under k independent hazards μ_i, $1 \leq i \leq k$, is assumed to be the minimum of the k lifetimes. Hence, the total force of mortality, $\mu(x)$, is

$$\mu(x) = \sum_{i=1}^{k} \mu_i(x) \quad .$$

Chiang assumes that the ratio

$$c_{it} = \frac{\mu_i(x)}{\mu(x)} \quad ,$$

over the time interval (x_t, x_{t+1}), is constant and independent of x. This condition, it is shown, can be satisfied by members of the three classes of extreme-value distributions. If the distribution must be of the same functional form, then these are the only permissible forms.

In another paper entitled "An Analysis of Some Failure Data," D. J. Davis [1952] presents failure data from 30 systems which are compared to exponential and normal distributions. Chi-square comparison is made on time-to-failure data by choosing convenient intervals of time, estimating the mean by the sample mean, and calculating the expected number of occurrences in each interval. When data are the number of failures in equal time intervals, a comparison is made to the Poisson distribution. When only a few time intervals are available, an index of dispersion is calculated as $\Sigma (X_i - \bar{X})/\bar{X}$. Normality of the data is assessed by visual observation and by a chi-square test on equal intervals of time. In the example considered, it is shown that the exponential distribution may be regarded as a useful approximation to the actual failure distribution.

Dixon, et al, present a study of the "Use of Triethylenethiophosphoramide (thio TEPA) as an Adjuvant to the Surgical Treatment of Gastric and Colorectal Carcinoma: Ten-Year Follow-Up." Data on survival from a colorectal study and gastric study are presented. For the colorectal study, the survival rate is fitted by the equation EXP (-bt), excluding 30-day deaths. For the gastric study, survival is fitted by EXP (-0.087t) + (1 - a) EXP (-bt)

using a nonlinear least square fit. The curves are described by the half-life, which is the number of years after the operation at which point 50 percent of the patients have died. In the gastric study, the proportion of patients with a half-life of 8 years is found and the half-life of the remaining cases is given. The half-life of 8 years was apparently found by trial and error. The survival rate is compared with the survival for the U.S. population of the same age, sex and race constitution, although neither the age of the patients in the study is indicated, nor was it used to adjust the survival rate.

In their paper "The Relative Survival Rate: A Statistical Methodology," Ederer, et al [undated] define the survival rate as the proportion of patients surviving a specified time interval at a given age. The relative survival rate, then, is the ratio of observed survival rate to the expected survival rate of a group of patients with similar characteristics of age, sex and race, but free of the specific disease under study. This rate is preferred over the computation of survival rate by excluding deaths from other causes then the disease under study. The relative rate is also called the age-adjusted survival rate, but it does not adjust for any association between age and risk of dying from the specific cancer under study. Life tables of the entire population may be used for the expected rate in most cases, since excluding the specific disease has a negligible effect on the overall rate. It is, however, pointed out that there may be a large effect, depending on the effect of associations such as that between smoking and

lung cancer. In making the adjustment, care must be exercised in selecting either the static or fluent life table. For analysis over an extended number of follow-up years, the varying age composition of the patient group must be considered in determining the expected survival rate.

An exact method, approximate methods, and a simple approximate method for calculating expected survival rates are given. The approximate methods divide the population into varying age groups. The exact method uses the survival probability for each individual and averages over the group.

The standard error of the relative rate is shown to be approximately equal to the standard error of the observed survival rate divided by the expected survival rate.

Epstein and Sobel [1953] in their article, "Life Testing," have shown that times-to-death are assumed to have an exponential distribution with mean θ. The estimator

$$\hat{\theta}_{r,n} = \frac{x_1 + x_2 + \cdots + x_r + (n-r)x_r}{r} ,$$

based on the first r out of n possible failures, is shown to be the maximum likelihood estimate, unbiased, minimum variance, efficient and sufficient, with density

$$\frac{1}{(r-1)} \left(\frac{r}{\theta}\right)^r y^{r-1} e^{-rg/\theta} \quad \text{for} \quad y > 0 .$$

Fix and Neyman [1951] in their paper entitled "A Simple Stochastic Model of Recovery, Relapse, Death and Loss of Patients," have found that cancer studies on either the same treatment applied to different categories of patients or of different treatments applied to a specific category of patients, usually must be based on small numbers of observations. The first step towards the development of statistical tests must consist in building a stochastic model of the phenomena studied. The model must involve simplifications. The following four states are defined:

$\mathcal{E}_0$ = initial state;

S_1 = dead from cancer or operative death;

S_2 = recovered from cancer or leading normal life; and

S_3 = lost after recovery, either by death due to other causes or inability to trace the patient.

A state diagram showing the possible transitions is given. The probability of transition from state i to j in the time interval k, $Q_{ij,k}(t_1, t_2)$, when in state i at time t_1 and state j at time t_2 is defined. The rate intensities

$$q_{ij} = \frac{\partial Q_{ij,k}(t_1, t_2)}{\partial t_2} \quad \text{for} \quad k = 0 \text{ or } 1$$

are defined and only one transfer in the time interval is assumed. The probabilities are assumed constant in small time intervals. The solutions for $Q_{ij}(t)$ where $t = t_2 - t_1$ are obtained and the estimation problem is

discussed. The notion of the expected length of normal life, state S_2, following a given treatment is introduced and calculated. A numerical calculation shows that different intensities can produce essentially the same death rates, but different normal life lengths.

A study of the "Maximum-Likelihood Estimation of the Parameters of the Gompertz Survival Function," is undertaken by Garg, et al [1970]. They define the Gompertz distribution survivor function as

$$\exp\left(-\frac{k}{\alpha}\left(e^{\alpha t}-1\right)\right)$$

and hazard rate as

$$\lambda(t) = k\, e^{\alpha t} \quad .$$

A shift in the time scale keeps the same functional form, but with a new constant for k.

The maximum likelihood equation for grouped data is given and its solution obtained by an iterative method. The empirical covariance matrix of the maximum likelihood estimates is derived. The estimation is applied to grouped data from a study on the mortality of mice, used to investigate the effect of oral contraceptives. A chi-square value is given, but it is not clear what was tested; it is stated that the distribution described the mortality quite well but no justification is given. The (MLE) maximum likelihood estimation appears to give a better fit than least square values.

The use and advantage of using the hazard rate is described by Gehan [1969] in his paper "Estimating Survival Functions from the Life Table." The author shows that the log normal survivorship function, although used in medical follow-up studies, has a hazard rate that is inappropriate for most medical situations. Methods of estimating the hazard rates, based on a partition of the time scale, are reviewed. Approximations to the variance of the estimates of the hazard rate and probability density functions are derived. A new statistic, the median remaining life-time, is proposed and an approximation to its variance is obtained.

In developing "An Approximating Function for the Hazard Rate," Ghare and Kim [1970] define the hazard rate as:

$$h(t) = \frac{a}{t^2 + b} + c\, e^{t(d/e)}, \qquad 0 < t$$

with $a,d \geq 0$ and $b,c,e > 0$. This hazard rate is used to approximate the bath-tub hazard shape. The parameters are estimated by a non-linear least squares fit of $h(t)$ with the empirical hazard rate determined by an arbitrary partition of the time period. Two examples are given, the first based on 1,000 failure times of B-52 aircrafts and the second based on time-to-failure of 480 fuses. Neither example, however, exhibited a bath-tub hazard rate.

Grenander [1956], in his paper "On the Theory of Mortality Measurement," discusses some of the problems encountered in this type of study.

Among others, the more basic problems that arise in biometric mortality studies are the small size of the sample and also that of choosing the best method, among the various methods available, for measuring mortality. The author suggests choosing a computationally simple method which has a relatively high efficiency, as measured by its mean square error.

Denote by ℓ_x, the probability at the time of birth for a person belonging to the population under study, of still being alive at age x. ℓ_x is normed so that $\ell_0 = 1$. The mortality intensity μ_x (also, otherwise known as the conditional failure rate, failure rate, or hazard rate) is then defined, as usual, by

$$\mu_x = -\frac{\ell'_x}{\ell_x}$$

where ℓ' denotes the derivative of ℓ with respect to x.

In Chapter 1 of Grenander's paper, he discusses some of the usual methods (maximum likelihood, method of moments and its variations and the King-Hardy method) for estimation of the so-called Makeham model where μ_x is given by

$$\mu_x = \alpha + \beta e^{\gamma x} .$$

Some tests for differential mortality (i.e., whether or not the mortality rates in two populations are the same against appropriate alternatives) are

also considered in this chapter.

In Chapter 2, the author deals with mortality measurements when it is not known and, therefore, do not assume any particular analytic form for the mortality intensity μ_x. Estimates and their properties are studied for the probabilities of death, reverses, etc.

The situation, when we have some information about the general form of the mortality curve, although we cannot express this knowledge as a parametric representation for μ_x, is considered in Chapter 3. For example, when the frequency function is known to be nonincreasing, it is shown that the maximum likelihood estimate is given by the largest convex minorant. A similar result was shown to hold for a nondecreasing mortality intensity.

Finally, in Chapter 4, the results of numerical computations on data of the efficiencies of various methods of estimation are presented. The question was left open to further research as to how to make a compromise between efficiency and the specific case of computation.

"Maximum Likelihood Estimation in Truncated Samples," is the subject of Halperin's [1952] paper. The maximum likelihood estimation of parameters from a censored sample of r out of n ordered observations is considered. The following regularity assumptions are required for the main theorem, which states that maximum likelihood estimators from the joint density function behave the same as those from identically distributed and independent samples.

The following Regularity Assumptions are made on Density $f(x,\theta_1,\ldots,\theta_p)$:

A. $\dfrac{\partial \ln f(x,\theta_1,\ldots,\theta_p)}{\partial \theta_i}$, $\dfrac{\partial^2 \ln f(x,\theta_1,\ldots,\theta_p)}{\partial \theta_i^2}$ and $\dfrac{\partial^3 \ln f(x,\theta_1,\ldots,\theta_p)}{\partial \theta_i^3}$

exist a.e.

B. $\left| \dfrac{\partial f}{\partial \theta_i} \right| < F_{1i}(x)$, $\left| \dfrac{\partial^2 f}{\partial \theta_i^2} \right| < F_{2i}(x)$,

$\left| \dfrac{\partial^3 f}{\partial \theta_i^3} \right| < F_{3i}(x)$, $\left| \dfrac{\partial^3 \ln F(x,\theta_1,\ldots,\theta_p)}{\partial \theta_i^3} \right| < H_i(x)$,

where $F_{1i}(x)$, $F_{2i}(x)$, $F_{3i}(x)$ are integrable over $(-\infty,\infty)$ and

$$\int_{-\infty}^{\infty} H_i(x)\, f(x,\theta_1,\ldots,\theta_p)dx < M_i$$

and M_i is independent of $\theta_1,\ldots,\theta_p$.

C. The matrix with elements A_{ij}

$$A_{ij} = \int_0^{\lambda} \left(\frac{\partial \ln f}{\partial \theta_i} \right)\left(\frac{\partial \ln f}{\partial \theta_j} \right) f\left(x,\theta_1,\ldots,\theta_p \right) dx$$

$$+ \frac{1}{p}\left(\int_0^{\lambda} \frac{\partial f(x,\theta_1,\ldots,\theta_p)}{\partial \theta_i}\, dx \right)\left(\int_0^{\lambda} \frac{\partial f}{\partial \theta_j}\, dx \right)$$

is positive definite for all $\theta_1,\ldots,\theta_p$, where λ is defined by

$$q = \int_{-\infty}^{-\lambda} f(x,\theta_1,\ldots,\theta_p)dx$$

and q is such that $\mu = [qn]$.

D. $f(x,\theta_1,\ldots,\theta_p)$ is continuous at $\lambda = x$ and its derivative is continuous with

$$\frac{\partial \ln f(x)}{\partial \theta_i} \quad , \quad \frac{\partial^2 \ln f(x)}{\partial \theta_i^2} \quad , \quad \frac{\partial^3 \ln f(x)}{\partial \theta_i^3} \quad ,$$

continuous at $x = \lambda$.

E. $1/(F(x,\theta_1,\ldots,\theta_p))$ is bounded independently of θ_i, $1 \le i \le p$, by $I_i(x)$ and $E(I_i(x,\theta_1,\ldots,\theta_p)) < I_i$ when I is independent of θ_i. Under these regularity assumptions, Halperin proves:

THEOREM: <u>The likelihood equations</u>

$$\frac{\partial \ln h(x_1,\ldots,x_r,\theta_1,\ldots,\theta_p)}{\partial \theta_i} = 0, \quad 1 \le i \le p,$$

<u>have roots</u> $\hat{\theta}_1,\ldots,\hat{\theta}_p$ <u>which</u>

1) <u>converge in probability to</u> $\theta_1,\ldots,\theta_p$;

2) <u>the</u> $\sqrt{n}\,(\hat{\theta} - \theta)$ <u>when</u> $\theta = (\theta_1,\ldots,\theta_p)$ <u>and</u> $\theta = (\hat{\theta}_1,\ldots,\hat{\theta}_p)$ <u>has</u>

a multivariate normal limit law; and

3) the concentration ellipsoid, corresponding to the covariance matrix of the limiting distribution of $\sqrt{n}\,(\hat{\theta} - \theta)$, is identical with

$$\sum_{i,j=1}^{p} E\left(\frac{\partial \log h}{\partial \theta_i} \frac{\partial \log h}{\partial \theta_j} \right)\left(\mu_i - \theta_i \right)\left(\mu_j - \theta_j \right) = p + 2 \ .$$

A study of the "Timing of the Distribution of Events Between Observations," is done in Harris, et al's paper [1950]. They have determined that the rate at which a disease occurs (the example considered herein is tuberculosis) can be estimated from knowledge of the time of last diagnosis and the time of first diagnosis. Lost cases are accounted for. Two cases are considered: the constant rate ρ, and stepped rates ρ_k. Letting n be the date of loss, that is, n is the last date before occurrence, and $n + m$ be the first date after occurrence, then a first approximation to the rate is given by

$$\frac{\Sigma' 1}{\Sigma\left(n + \frac{m}{2} \right)}$$

where Σ indicates the summation over all cases and Σ' indicates summation over eventful cases. In the stepped case

$$\rho_k = \frac{\Sigma f_k}{\Sigma\left(n_k + \frac{1}{2} m_k\right)}$$

where f_k depends on the estimated ρ_k,

$$f_k = \frac{\rho_k m_k}{\Sigma_k \rho_k m_k} .$$

Kale [19 2] presented a paper on "The Solution of Likelihood Equations by Iteration Processes. The Multiparametric Case." The metric

$$b(A) = \max_{ij} |\lambda_{ij}|$$

for a matrix $A = (\lambda_{ij})$, $1 \leq i \leq k$ and $1 \leq j \leq k$, is used to prove that as the sample size approaches infinity and under regularity conditions on the density, the iterative methods of solving maximum likelihood equations converge. A consistent estimate of the parameters must be used as the starting values. The general iterative method is

$$T_{r+1} = \Phi(T_r)$$

where

$$\Phi(T_r) = T_r - G(T_r)\, \Lambda(T_r)$$

and

$$G(\theta) = \left(\frac{\partial \log L}{\partial \theta_1}, \ldots, \frac{\partial \log L}{\partial \theta_k} \right)$$

and Λ is a matrix. It is shown that if

$$b\left[\frac{\partial \theta}{\partial \hat{\theta}}\right] \leq \frac{\alpha}{k}, \quad 0 \leq \alpha < 1$$

and $T_1 \in N_\rho(\hat{\theta})$ where $N_\rho(\hat{\theta})$ is a ρ-neighborhood of $\hat{\theta}$, then the iterative method converges for large sample sizes. Both the Newton-Raphson method which uses

$$\Lambda(\theta) = \left[\frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \right]^{-1}$$

and the method of scoring which assumes that

$$\Lambda(\theta) = -\frac{1}{N} [J^{rs}(\theta)]$$

where J is the information matrix are shown to satisfy the conditions.

Kimball [1960], using animals in his experiments reviews estimation procedures for mortality intensities, μ_x. He considers four methods, each using various rules for partitioning the time interval. Two methods of sampling are discussed:

1) the time intervals are chosen, usually of equal width, and the

number of deaths in the i^{th} interval, y_i, are recorded; and

2) the y_i are preassigned and the boundaries of the intervals, x_i to x_{i+1}, are observed.

The classical estimate of mortality at the midpoint of the i^{th} interval is

$$\mu^{(1)}_{x_i} + \frac{1}{2} T_i = \frac{y_i / T_i}{\frac{1}{2}(N_i + N_{i+1})}$$

where $T_i = x_{i+1} - x_i$ is the interval width and N_i is the number alive at time x_i. Another estimate is

$$\mu^{(2)}_{x_i} + \frac{1}{2} T_i = \frac{1}{T_i} \log \frac{N_i}{N_{i+1}} .$$

A numerical study of these estimates showed the bias to be a function of all the parameters, assuming a Gompertz mortality function.

For the second type of sampling, the estimate used is

$$\mu^{(3)}_i = \frac{y_{i-1}}{\sum_{j=0}^{y_i=1} (N_{i-j})(t_{i,j+1} - t_{ij})} ,$$

and it is unbiased for μ_i if $\mu_{x_i} + \tau = \mu_i$ for $0 \leq i < T$, i.e., constant mortality in the interval. If μ_i increases linearly over the interval,

the estimate

$$\mu_{i+1}^{(4)} = \frac{(y_i - 1) - A_i \mu_i^{(4)}}{B_i}$$

where A_i and B_i are complicated functions of the statistics is proposed.

The four estimates are applied to statistics on time-to-death of rats exposed to radiation. Differences of up to 8 percent between the four estimates are found. Comparison due to different partitions is not discussed. Plots of $\mu^{(3)}$ and $\mu^{(4)}$ are given and show several dips.

In his paper, "A New Response Time Distribution," Kodlin [1967] gives an increasing mortality rate, $\mu(t)$ as

$$\lambda(t) = c + kt$$

with probability density function, $f(t)$, as

$$f(t) = (c + kt)e^{-(ct+\frac{1}{2}kt^2)}, \qquad t,c,k \geq 0 ,$$

for the time-to-death of an individual under a specific treatment. The mean and variance are obtained. The maximum likelihood equations are given for grouped data with censored trails, using: n = number of individuals in a sample; d_i = number dying at t_i; s_i = number alive at τ_i; and, τ_i = time of loss or end of follow-up.

The model is applied to 121 breast cancer patients to compute the excess risk by using only those that die from causes other than cancer. The overall mortality is decreasing and cannot be used with the model.

A second model for mortality, $\lambda(t)$, is proposed

$$\lambda(t) = \frac{\delta\lambda e^{-\lambda t}}{\delta e^{-\lambda t} + 1 - \delta} + c + kt$$

as are some "crude" estimators.

Interpretation of $f(t)$ for various parameters is given, as applied to medical follow-up studies.

Mantel [1966] proposes the overall comparison of two survival functions, requiring a value function for rating the duration of survival. A chi-square procedure is proposed with an implicit value function. The study begins with a population of N_{1l} homogeneous patients. In the i^{th} interval of the study, r_{1j} die and ℓ_{1i} are lost track of. The estimated probability of surviving to the j^{th} period, then, is

$$\sum_{i=1}^{j/1l} \frac{N_{1i} - r_{1i}}{N_{1i}}$$

where $N_{1i+1} = N_{1i} - r_{1i} - \ell_{1i}$.

A chi-square test, based on a contingency table is proposed.

In their paper, "Life Tests Under Competing Causes of Failure and the Theory of Competing Risks," Moeschberger, et al [1971] use the nonnegative random variables, y_ℓ, to denote the lifetime of a subject due to cause-of-failure, C_ℓ, $\ell = 1,\ldots,k$. It is assumed that $y_1 = \min_\ell y_\ell$ and C_ℓ are observed. Parametric distributions are assumed, in particular, a Weibull distribution with ungrouped data are desired. Results with this method are smoother than similar tests made using nonparametric or partially nonparametric methods.

Murthy [1965a] in a paper entitled "Estimation of Probability Density," defines a function $k(x)$ as a window, if

1) $k(x) \geq 0$;

2) $k(x) = k(-x)$;

3) $\lim_{|x| \to \infty} x\, k(x) = 0$; and

4) $\int_{-\infty}^{\infty} k(x)dx = 1$.

It is shown that the estimate

$$f_n(t) = \int_{-\infty}^{\infty} B_n\, k(B_n(x-t))\, dF_n(x)$$

where $F_n(x)$ is the empirical distribution and B_n is a sequence of nonnegative constants with $B_n \to \infty$ and $B_n/n \to \infty$ as $n \to \infty$ is consistent at a point of continuity of $f(x)$.

In another paper reviewing the "Estimation of Jumps, Reliability and

Hazard Rate," Murthy [1965b] defines the reliability estimator

$$R_n^*(t) = \frac{B_n}{n} \sum_{j=1}^{n} \int_t^{\infty} k(B_n(x_j - x))dx$$

as having the large sample mean, $R(t)$, and variance $R(t)(1 - R(t))$, which is asymptotically normal. The hazard rate estimator is defined as

$$Z_n(t) = \frac{f_n(t)}{R_n(t)}$$

where $R_n(t) = \frac{1}{n}$ times the number of observations greater than t among $x_1,\ldots,x_n$, and where $f_n(t)$ is as defined by Murthy in [1965a] and is shown for large samples to be consistent. The estimator

$$Z_n^*(t) = \frac{f_n(t)}{R_n^*(t)}$$

is shown to be consistent and asymptotically normal.

Murthy and Swartz [1971] in their report "Contributions to Reliability," present various realistic models for life distributions in Chapter IV that correspond to density functions and that can be expressed in closed form. Some methods of estimation, using order statistics, are also suggested.

Two factors are considered in Sampford's [1952] paper on the estimation of response-time distributions. They are:

I. fundamental concepts and general methods, and

II. multi-stimulus distributions.

A normal distribution of the response time is assumed and the experiment is carried through to completion. Maximum likelihood methods are developed for grouped data. A second part, considers estimation of parameters in the distribution of response time when two causes of death are operating on the subject. The main model assumes that the joint action of the two causes of death is independent and the density of the primary cause is normal. Various densities for the secondary cause of death are considered, including those of the exponential and normal types.

Swartz, in his paper "The Mean Residual Lifetime Function," [1973] establishes conditions and properties on a function which characterize that function as a mean residual lifetime function.

The elementary properties of hazard functions are discussed by Watson and Leadbetter in "Hazard Analysis, I" [1964]. Estimation of hazard rates using windows is presented with the estimator being

$$h_n^*(\gamma) = \sum_{r=1}^{n} \frac{\delta_n(x - X_{(r)})}{n - r + 1}$$

and

$$E[h_n^*(x)] = \int_0^\infty \delta_n(x-y)\, h(y)dy - \int_0^\infty \delta_n(x-y)\, F^n(y)\, h(y)dy \quad ,$$

with an expression for Var (h_n^*) given also. δ_n is a window function. Some less sophisticated estimators are given: a graphical derivative method, a histogram estimator, and the classical actuarial estimator. Death dates on mice, exposed to gamma radiation, are used to obtain hazard estimates by various methods.

Zelen [1966] explores the application of statistics and the role it can play by the increased use of probability models to guide the researcher in the types of data to collect. Suggestions for further experiments and as a frame of reference are discussed. The time-to-failure is assumed to have a survival function

$$S(t) = \begin{cases} e^{-\lambda(t-G)}, & t > G \\ 1, & t \leq G \end{cases},$$

where G is the guarantee time. The mean is $G + (1/\lambda)$. This model is applied to animal tumor systems and acute leukemia, destruction of tumor cells with laser energy and the analysis of survival data, with concomitant information. In the last case, it is assumed that time-to-failure has density

$$f_i(t) = \lambda_i e^{-\lambda_i t} \quad \text{for all} \quad t \geq 0$$

for the i^{th} patient, and that $1/\lambda_i = a + bx_i$ where x_i is the

concomitant variable, such as the logarithm of white blood count, as in the sample subjects with acute myelogenous leukemia.

CHAPTER III

Realistic Models for Mortality Rates and Their Estimation

Introduction

In this chapter mathematical models are advanced which, as special cases, represent constant, increasing and decreasing mortality rates, along with combinations of these properties. Bath-tub shaped mortality rate curves are the general shape of these models. The first part of the tub corresponds to infantile mortality, the second part (more or less, constant mortality) corresponds to useful life, and finally, the last part, which is increasing corresponds to decay, aging, etc., culminating in death. Their corresponding probability distributions and survivorship functions are obtained in closed form. To the best of our knowledge, the closed form representations of survivorship functions corresponding to the wide class of forces of mortality, considered herein are new to the literature relevant to mortality rates and their estimation. The variety of mortality shapes derived are illustrated and applied. Estimation is performed by several methods: empirical, various window-smoothing, and maximum likelihood techniques.

Estimation of the Hazard Rate Function

Observed times-to-failure can be used to estimate the hazard rate function (mortality rate or risk will also be used, interchangeably). Non-parametric methods for estimating hazard rates were considered by Murthy [1968] and Grenander [1956]. However, the most they can do is to throw light on the shape of the underlying hazard rate function and cannot be used for decision-making and prediction. Thus, parametric procedures based on realistic models

for the hazard rate function are the only suitable approaches for estimation and prediction. It is generally accepted that the most general hazard rate curve consists of an early decreasing phase, followed by a section of constant failure (corresponding to the useful life period), followed by a rapidly increasing failure rate as death approaches.

Theory

Let the nonnegative random variable, T, be the time to death of the subject. Let T have a cumulative distribution function (cdf)

$$F(t) = P(T < t) \tag{1}$$

and assume that T has a mean, μ, variance, σ^2, and a probability density function

$$f(t)dt = dF(t) \ . \tag{2}$$

The mortality rate, $\lambda(t)$, is the instantaneous chance of dying at time t, given that the subject has survived to time t and is defined as

$$\lambda(t) = \frac{f(t)}{R(t)} , \qquad t \geq 0 \tag{3}$$

where

$$R(t) = 1 - F(t) \tag{4}$$

is the survivorship function. That $\lambda(t)$ is a mortality rate can be seen by finding the probability of death in the time interval $(t , t + \Delta t)$, where

Δt is a small positive time increment, given survival to time t. We have from the definition of conditional probability that

$$\frac{P(t \leq T < t + \frac{\Delta t}{T} > t)}{P(T \geq t)} = \frac{F(t + \Delta t) - F(t)}{R(t)} \quad . \tag{5}$$

By the mean value theorem of calculus, $F(t + \Delta t) - F(t) = f(\xi)\Delta t$, for some ξ between t and $t + \Delta t$, hence

$$P(t \leq T < t + \frac{\Delta t}{T} \geq t) = \frac{f(\xi)}{R(t)} \Delta t, \qquad t \leq \xi \leq t + \Delta t \quad . \tag{6}$$

In the limit, as $\Delta t \longrightarrow 0$, the conditional probability is $\lambda(t)dt$.

The mortality rate also uniquely defines the survivorship function, as can be seen by writing the difference equation for the probability of death by time $t + \Delta t$:

$$F(t + \Delta t) = F(t) + [1 - F(t)]\, \lambda(t)dt \quad . \tag{7}$$

The right-hand side is the sum of the probabilities of two mutually exclusive events,

1) death by time t with probability $F(t)$; and
2) no death until time t and death in the interval $(t, t + \Delta t)$, with probability $[1 - F(t)]\lambda(t)\Delta t$.

Dividing both sides by Δt and taking the limit as $\Delta t \longrightarrow 0$ gives

$$\frac{dF(t)}{dt} = [1 - F(t)]\,\lambda(t) \tag{8}$$

which can be written as

$$\lambda(t) = -\frac{d}{dt} \ln R(t) \quad . \tag{9}$$

Solving this differential equation for $R(t)$ and using the initial condition $R(0) = 1$ and $R(\infty) = 0$, we obtain

$$R(t) = e^{-\int_0^t \lambda(x)dx} \quad . \tag{10}$$

The integral $\Lambda(t) = \int_0^t \lambda(x)dx$ is also called the accumulated, or cumulative hazard by time t; its derivative $\lambda(t)$ is the hazard rate.

Another interpretation of the mortality rate is to compute the ratio of expected deaths in the interval $(t, t + \Delta t)$ to the expected number of subjects alive at time t. Let $n(t)$ be the expected number alive at time t from an original population of $N = n(0)$, then

$$n(t + \Delta t) = N[R(t + \Delta t) - R(t)] \quad ; \tag{11}$$

dividing both sides by Δt, and taking the limit as $\Delta t \longrightarrow 0$, we have

$$\frac{dn(t)}{dt} = -Nf(t) = -N \frac{f(t)}{1 - F(t)} [1 - F(t)] = -n(t)\,\lambda(t) \tag{12}$$

or

$$\lambda(t) = -\frac{1}{n(t)} \frac{dn(t)}{dt} \quad . \tag{13}$$

This differential equation may be solved for $n(t)$, producing

$$n(t) = N e^{-\int_0^t \lambda(x)dx} . \tag{14}$$

The mortality rate is sometimes referred to as a conditional density; however, $\lambda(t)$ is not a density since $\int_0^\infty \lambda(x)dx = \infty$. $\lambda(t)$ is, in fact, $f(t/(T \geq t))$ which is a conditional density only when $x = t$, but is not a conditional density as t varies.

We are concerned in this chapter with selecting a mortality function that models a medical follow-up study and that also can be integrated to give a survivorship function in closed form. Murthy and Swartz [1971] show that any function, $h(x)$, which is positive and for which

$$\lim_{t \to \infty} \int_0^t h(x)dx \tag{15}$$

is unbounded, can be used as a mortality rate which will give a unique survivorship function

$$R(t) = e^{-\int_0^t h(x)dx} . \tag{16}$$

Three models are discussed in the following section.

Realistic Models of the Mortality Rate

MODEL I: This class of distributions is characterized by the hazard

rate

$$\lambda(t) = -\frac{1}{1+bt} + c\,dt^{d-1}, \qquad t \geq 0 \tag{17}$$

where the parameters, $a,b,c \geq 0$ and $d \geq 1$. Some special cases of this model are: when $c = 0$,

$$\lambda = \frac{a}{1+bt} \quad , \text{ a decreasing function of time} \quad . \tag{18}$$

The mortality rate reduces to a constant when $a = 0$, $d = 1$, or $c = 0$, $b = 0$ and we have an exponential distribution. When $b = 0$, $d = 2$, we have the linear rate used by Kodlin [1967] to model death from breat cancer. Finally,

$$\lambda(t) = c\,dt^{d-1} \quad \text{when} \quad a = 0 \quad , \tag{19}$$

corresponds to a two-parameter Weibull distribution.

The cumulative distribution function for Model I is

$$F(t) = 1 - \frac{1}{(1+bx)^{a/b}}\, e^{-ct^d} \tag{20}$$

and the density function is

$$f(t) = \left(\frac{a}{1+bt} + c\,dt^{d-1}\right)\left(\frac{1}{(1+bt)^{a/b}}\, e^{ct^d}\right) \quad . \tag{21}$$

MODEL II: This class of distributions has the hazard rate

$$\lambda(t) = ab\, e^{-bt} + c\, dt^{d-1}; \qquad d \geq 1,\ a,b,c > 0 \ . \tag{22}$$

The corresponding cumulative distribution function is

$$F(t) = 1 - e^{-a(1-e^{-bt})-ct^d} \tag{23}$$

and the density is

$$f(t) = (ab\, e^{-bt} + c\, dt^{d-1})\, e^{-a(1-e^{-bt})-ct^d} \tag{24}$$

The special case of (24), where $a = 1/b^2$, may be of interest since it is a three parameter family of distributions with the bath-tub property. This class of distribution is given by

$$F(t) = 1 - e^{-\frac{1}{b^t}(1-e^{-bt})-ct^d} \tag{25}$$

with density function

$$f(t) = \left(\frac{e^{-bt}}{b} + c\, dt^{d-1} \right)\left(e^{b^{\frac{1}{2}}(1-e^{-bt})-ct^d} \right) \ . \tag{26}$$

Note that the two-parameter Weibull and Kodlin's models are also special cases of this general model.

MODEL III: This class has the hazard rate

$$\lambda(t) = \frac{a}{1 + bt} + cde^{dt}; \qquad a,b,c,d \geq 0 \ . \tag{27}$$

The corresponding cdf and density are, respectively, given by

$$F(t) = 1 - \frac{1}{(1 + bt)^{a/b}} e^{-c(e^{dt}-1)} \tag{28}$$

and

$$f(t) = \left(\frac{a}{1 + bt} + cde^{dt} \right)\left(\frac{1}{(1 + bt)^{a/b}} e^{-c(e^{dt}-1)} \right) \ . \tag{29}$$

When $a = 0$ and $d > 0$, $\lambda(t) = cde^{dt}$, which is a monotonically increasing hazard rate called the Gompertz rate by Broadbent [1958] and applied to the lifetimes of milk bottles. This rate is studied by Garg, et al [1970] and is used to model mortality of mice in oral contraceptive studies.

Application to Medical Follow-Up Studies

The mortality rates corresponding to many real life situations fall within the domain of our models. From the practical point of view, we distinguish among three categories of mortality rates: BTR, DHR and IHR, depending on their respective shape. Category BTR is marked by a decreasing risk of mortality function, followed by an increasing one. Category DHR is purely decreasing, while that of IHR is increasing. The classical example of

a bath-tub hazard curve is the survival data of humans (as well as other animals). It is characterized by a high infant mortality during the first year of life, slowly decreasing risk from ages 1 through 4, a low mortality rate from ages 5 to around 35, and increasing risk for older ages, as shown in Figure 1.

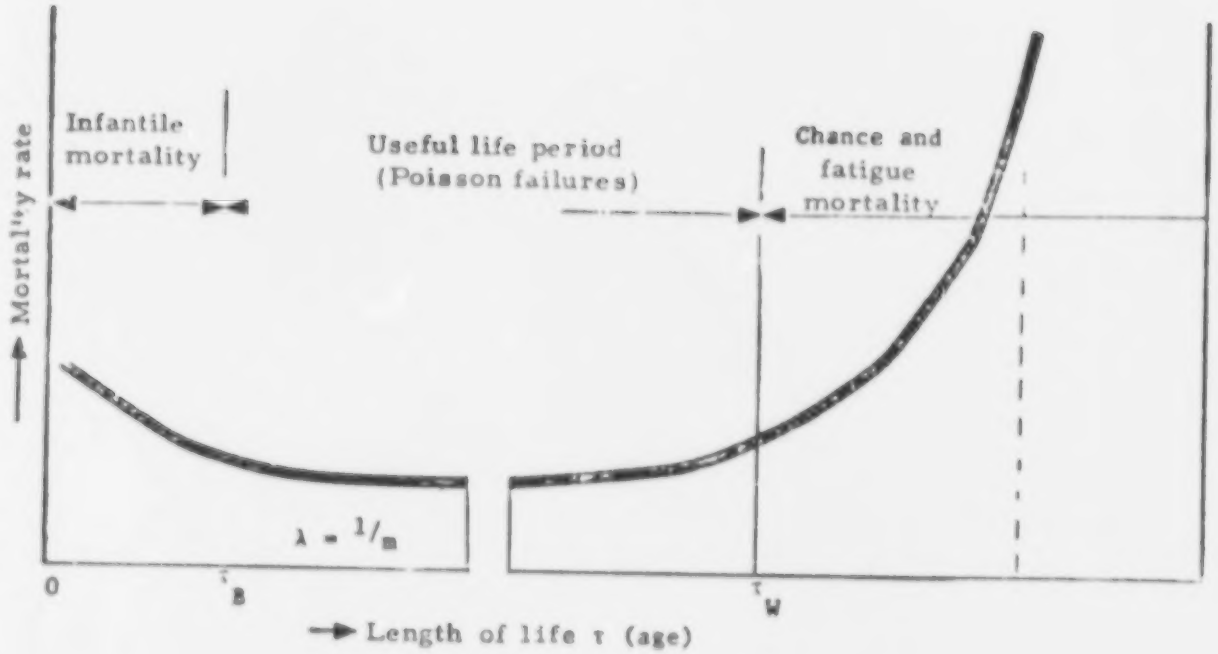
Usually in statistical analysis of data a cumulative distribution function is assumed. The random variable under study, in this case, the time-to-death, can in many situations be described by a mathematical model which also determines the functional form of the distribution function. Zelen [1966] has given several such models for death due to specific types of cancer. These models lead him to consider the exponential distribution. Davis [1952] showed that the exponential model gave a good fit in the case of thirty nonhuman systems. Unfortunately, the exponential distribution has a constant mortality rate which may be a reasonable assumption in the case of certain short periods of time. That it is not the case for time periods over five years, in medical follow-up studies on humans, was shown by Berg and Robbins [1967]. In the absence of a mathematical model to guide us in the selection of a distribution, the concept of the mortality rate has been adopted. Buckland [1964] discusses many examples of hazard rates that correspond to the Weibull, gamma, normal, log normal, and extreme value distributions. Boag [1949] considered the Weibull, gamma, log normal and exponential distributions in a study on deaths from cancer after treatment. He found that the best fit to be given by the log normal distribution; and, for

a large group, suggested that the survivorship be modeled by a mixture of c percent completely cured and $(1-c)$ percent whose survival times follow a log normal distribution. Gehan [1969] showed that the mortality rate of the log normal distribution was inappropriate for medical follow-up studies and that a bath-tub rate was necessary.

The DHR class of distributions characterize situations where there is an initial high risk which decreases with time. Over a limited time period, this might describe survival data of patients after transplants or implantations of mechanical devices, besides the usual infantile mortality in a "normal" situation.

The IHR class of distributions fit situations where a very high mortality rate is reduced and the patient is given a longer time to live, but is not "cured." This class may be a useful approximation to the mortality rate over the 100 odd years of human life, because the high mortality rate of the final five years is several orders of magnitude greater than the initially high infant death rate. Broadbent [1958], Garg, et al [1970], Kodlin [1967] and Grenander [1956] consider this type of hazard.

To demonstrate the versatility of the three models proposed in this chapter, with respect to their ability to describe the various mortality rates that one encounters in real life, an interactive graphics program was written so that a person at the console can select one of the models and the parameter values. The corresponding hazard function is displayed on the screen. If desired, increments of the parameters can be typed in, and plots



Typical mortality rate as a function of age

Figure 1.

of these curves displayed, one after the other, as the parameters increase or decrease. Thus, the characteristics of each model can be fully illustrated. Pictures of some of these plots are shown in Chapter 4 of this report.

Life Table Model

Assume a homogeneous collection of people. Let the time-to-death of a person in the group be a positive random variable, T, with hazard $Z(t)$ of either our model I, II or III. At a particular time, we have recorded n time-to-deaths, $t_1 \leq \cdots \leq t_n$, of the first n people to die. One problem which arises is to estimate the parameters in the model. Another is to compare the hazard of one such group with another.

Follow-Up Model

In this model, a fixed number of people, n, at the same age, all are assumed to have similar mortality rates when observed for a fixed length of time, L, and their times-to-death are recorded as, $t_1 \leq t_2 \leq \cdots \leq t_r$, where r is an integer-valued random variable $r \leq n$. (If $r < n$, we refer to the sample as being truncated.) The problem is to estimate the model parameters and to compare the hazard with other groups.

A probabilistic model of the preceding models is obtained by letting a positive random variable, T, be the time-to-death of an individual and assuming that the individual times-to-death are independent and identically distributed random variables with common survivorship, $R(t)$. A random sample

of size n of the random variable T will be denoted by $t_1,\ldots,t_n$. In the first model, the times are observed in order, and hence, correspond to the order statistics, $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$, with the distribution of $t_{(i)}$ being

$$P(t_{(i)} \leq t) = i\binom{n-1}{i-1} F^{i-1}(t) R^{n-i}(t), \qquad 0 \leq t \tag{30}$$

and with joint density

$$g(t_1,\ldots,t_\ell) = \begin{cases} n! f(t_1) \cdots f(t_\ell), & 0 \leq t_1 \leq t_2 \leq \cdots \leq t_\ell \\ 0, & \text{otherwise} \end{cases}. \tag{31}$$

Even when the sample is censored, information is provided by the non-failed items, as indicated by the $R^{n-i}(t_1)$ term.

Variation in these models, to consider patients lost from follow-up, and the possibility of contamination by an individual with a markedly different survivorship will be considered in a following section.

We will now consider several estimates of the hazard rate at the random points, $t_{(i)}$, where $t_{(i)}$ is the i^{th} order statistic. The three models will be fitted, by nonlinear least-square technique, to the estimated hazard.

Hazard rate, $\lambda(t)$, defined as

$$\lambda(t) = -\frac{1}{R(t)}\frac{dR(t)}{dt} = -\lim_{h \to 0} \frac{R(t+h) - R(t)}{hR(t)}, \tag{32}$$

suggests the estimator, $Z_n(t)$:

$$Z_n(t_i) = -\frac{\frac{n_i + h}{n_i} - \frac{n_i}{n}}{h\,\frac{n_i}{n}} = \frac{1}{h}\,\frac{m_i}{n_i} \qquad (33)$$

where $n_i = \text{No.}\{j : t_j > t_i\}$ is the number of persons alive at time, t_i, $n_{i+h} = \text{No.}\{j : t_j > t_i + h\}$ is the number alive at time, $t_i + h$, and $m_i = n_i - n_{i+h}$ is the number who die in the time interval (t_i, t_{i+h}). Because the ratio of random variables is involved, it has not been possible to evaluate the expected value of this estimator.

Another estimator is provided by the definition,

$$\lambda(t) = -\frac{d}{dt} \ln R(t) \quad , \qquad (34)$$

$$\lambda(t) = -\lim_{h \to 0} \frac{\ln R(t + h) - \ln R(t)}{h} \quad . \qquad (35)$$

This suggests the estimator, which we call the empirical estimator:

$$Z_n(t_i) = -\frac{\ln n_{i+h} - \ln n_i}{h} = -\frac{1}{h} \ln \frac{n_{i+h}}{n_i} \quad \text{for} \quad n_{i+h} > 0 \quad . \qquad (36)$$

For small m_i, This estimate approaches $Z_n(t_i)$ of formula (33). The following program segment will compute $Z_n(t_i)$ for time-to-death dates, in days, stored in order in array, T(I), with TOT = n_i and SUM = $n_i + h$.

```
      DO 100 I = 1,N
      TOT = 0
      SUM = 0
      DO 200 J = I,N
      IF(T(J).GT.T(I)) TOT = TOT +1
  200 IF(T(J).GT.(T(I)+H)) SUM = SUM+1
      IF(SUM.LE.0.) SUM = 1
  100 ZN(I) = ALOG(TOT/SUM)/H.
```

We compute the expected value of the empirical estimator as

$$E Z_n(t_i) = -\frac{1}{h} (E \ln n_{i+h} - E \ln n_i) \quad . \tag{37}$$

Since n_i has a binomial distribution

$$P(n_i = k) = \binom{n}{k} F^k(t_i) F^{n-k}(t_i), \tag{38}$$

we have

$$E Z_n(t_i, h) = -\frac{1}{h} \left[\sum_{k=0}^{n} \ln k \binom{n}{k} (G(t_i + h) - G(t_i)) \right] \tag{39}$$

where

$$G(t) = F^k(t) F^{n-k}(t) \quad . \tag{40}$$

The expected value, as h is made arbitrarily small, is

$$E Z_n(t_i) = \lim_{h \to 0} E Z_n(t_i, h) = -\sum_{k=0}^{n} \ln k \binom{n}{k} G'(t_i) \quad . \tag{41}$$

Evaluating $G'(t)$, we obtain

$$E\,Z_n(t_i) = \frac{f(t_i)}{R(t_i)}\, b(n, t_i, \theta_1, \ldots, \theta_p) \quad . \tag{42}$$

where $b(\cdot)$ is a bias factor that depends on n, t_i and the parameters $\theta_1,\ldots,\theta_p$ of the survival and is expressed as:

$$b(n, t_i, \theta_1, \ldots, \theta_p) = \sum_{k=0}^{n} \ln k \binom{n}{k} R^k(t_i)(1 - R(t_i))^{n-k} (k - nR(t_i)) \quad . \tag{43}$$

If the time-to-death has an exponential survival, $R(t) = e^{-t/\theta}$, $t \geq 0$, it was shown by Epstein and Sobel [1953] that a minimum variance unbiased estimator of θ is $\tilde{\theta}$, where

$$\tilde{\theta} = \frac{T_{(1)} + \cdots + T_{(i)} + (n-i)T_{(i)}}{i} \quad . \tag{44}$$

Since the hazard $\lambda(t) = 1/\theta$, we propose the estimator $Z_n(t_i)$,

$$Z_n(t_i) = \frac{1}{\tilde{\theta}} \quad . \tag{45}$$

Preliminary investigation of this estimate, under different survivorships, is not encouraging as the estimator exhibits severe bias.

Estimation of densities by using "windows" has been applied to hazard rate estimation by Murthy [1965]. A window $K(w)$ is a function satisfying the conditions

1. $K(w) = K(-w)$;

2. $K(w) \geq 0$;

3. $\int_{-\infty}^{\infty} K(w) = 1$; and

4. $\lim_{|w| \to \infty} wK(w) = 0$.

Murthy [1965] shows that the estimator, called the window estimator,

$$Z_n(t_i) = \frac{f_n^*(t_i)}{R_n^*(t_i)} \tag{46}$$

is a consistent estimator for $\lambda(t)$, where

$$f_n^*(t) = \frac{\ln n}{n} \sum_{j=1}^{n} K(\ln n(t_j - t_i)) \tag{47}$$

and

$$R_n^*(t) = \int_t^{\infty} f_n^*(x)dx \quad . \tag{48}$$

For a rectangular window

$$K(w) = \begin{cases} 1/2h , & -h \leq w \leq h \\ 0 , & \text{otherwise} \end{cases} , \tag{49}$$

we have the <u>rectangular window</u> estimator

$$Z_n(t) = \frac{\frac{1}{n\,2h_1}\ \text{No.}\{j : t - h_1 \le t_j \le t + h_1\}}{\frac{1}{n}\left[\text{No.}\{j : t_j \ge t - h_1\} + \sum_{j \in A}\left(\frac{(t_j - t)}{2h} + \frac{1}{2}\right)\right]} \tag{50}$$

where Set $A = \{j : t - h_1 \le t_j \le t + h_1\}$, and $h_1 = h/(\ln n)$.

The expression is derived in the following way. Let

$$D(w) = \ln nK(\ln n - w) \ .$$

Then

$$D(w) = \begin{cases} 1/2h_1, & -h_1 \le w \le h_1 \\ 0, & \text{otherwise} \end{cases} , \tag{51}$$

hence,

$$f_n^*(t) = \frac{1}{n} \sum_{j=1}^{n} D(t_j - t) \ . \tag{52}$$

Since, $D(t_j - t)$ is $1/2h_1$ when $|t_j - t| \le h_1$, the sum counts the number of elements in the Set $A = \{j : t - h_1 \le t_j \le t + h_1\}$. The estimate of $R(t)$ given by $\int_t^\infty f_n^*(x)dx$ which, upon substituting $w = t_j - x$,

$$\int_t^\infty f_n^*(x)dx = \int_{-\infty}^{t_j - t} D(w)dw$$

$$= \begin{cases} 1, & t_j - t \ge h_1 \\ (t_j - t + h_1)/2h_1, & -h_1 \le t_j - t \le h_1 \\ 0, & \text{otherwise} \end{cases} , \tag{53}$$

we obtain the denominator of $Z_n(t)$.

A computational simplification of the window estimators was suggested by Watson and Leadbetter [1964]. The denominator of the window estimator was replaced by the estimator, $R_n(t_i) = 1 - i/(n+1)$. Using the density estimator $f_n^*(t)$ suggested by Murthy [1965], we consider the window estimator

$$Z_n(t_i) = \frac{f_n^*(t_i)}{\frac{n+1-i}{n+1}} \quad . \tag{54}$$

Since the denominator is not a random variable, we can write the expected value of this as,

$$\begin{aligned} E Z_n(t_i) &= \frac{n+1}{n+1-i} \ln n \, E \sum_{j=1}^{n} k(\ln n(T_j - t_i)) \\ &= \frac{n+1}{n+1-i} \ln n \sum_{j=1}^{n} \int_0^\infty K(\ln n(t - t_i)) \, f_j(t) dt \quad , \end{aligned} \tag{55}$$

where $f_j(t)$ is the density of the j^{th} order statistic

$$f_j(t) = j \binom{n-1}{j-1} F^{j-1}(t) \, R^{n-j-1}(t) \, f(t) \quad . \tag{56}$$

This is applied for rectangular and triangular windows to the three sets of model data.

A triangular window estimator is obtained when

$$K(w) = \begin{cases} 0 , & w < -h \\ \frac{w}{h^2} + \frac{1}{h} , & -h \le w \le 0^- \\ -\frac{w}{h^2} + \frac{1}{h} , & 0 < w \le h \\ 0 , & h < w \end{cases} . \tag{57}$$

Let $D(w) = \ln n K(w \ln n)$, then,

$$D(w) = \begin{cases} 0 , & w < h_1 \\ \frac{w}{h_1^2} + \frac{1}{h_1} , & -h_1 \le w \le 0 \\ -\frac{w}{h_1^2} + \frac{1}{h_1} , & 0 < w \le h_1 \\ 0 , & h_1 < w \end{cases} , \tag{58}$$

where $h_1 = h/(\ln n)$.

The density estimator is

$$f_n^*(t) = \sum_{j=1}^{n} D(t_j - t) \tag{59}$$

and

$$f_n^*(t_i) = \frac{1}{n h_1} \sum_{j \in A_i} \left(1 - \frac{t_j - t_i}{h_1} \right) \tag{60}$$

where the set $A_i = \{j : |t_j - t_i| \le h_1\}$.

The reliability estimator is

$$R_n^*(t) = \int_t^\infty f_n^*(x)dx = \frac{1}{n}\sum_{j=1}^{n}\int_t^\infty D(t_j - x)dx$$
$$= \frac{1}{n}\sum_{j=1}^{n}\int_{-\infty}^{t_j-t} D(u)du \quad , \tag{61}$$

where

$$\int_{-\infty}^{t_j-t} D(u)du = \begin{cases} 0 & , \quad t_j - t < -h_1 \\ \frac{(t_j-t)^2}{2h_1} + \frac{(t_j-t)}{h_1} + \frac{1}{2} & , \quad -h_1 \le t_j - t \le 0 \\ -\frac{(t_j-t)^2}{2h_1} + \frac{(t_j-t)}{h_1} + \frac{1}{2} & , \quad 0 < t_j - t \le 0 \\ 1 & , \quad h_1 < t_j - t \end{cases} \tag{62}$$

and, the hazard rate estimator is

$$Z_n(t_i) = \frac{\frac{1}{n h_1}\sum_{j\in A_i}\left(1 - \frac{t_j - t_i}{h_1}\right)}{\frac{1}{n}\left[\text{No.}\{j: t_j > t_i + h_1\} + \sum_{j\in A_i}\left(\frac{1}{2} + \frac{|t_j - t_i|}{h_1} + \text{SGN}(t_j - t_i)\frac{(t_j - t_i)^2}{2h_p}\right)\right]} \tag{63}$$

where $\text{SGN}(x) = x / |x|$, (i.e., the sign of x).

The simplified triangular-window estimator is

$$Z_n(t_i) = \frac{\frac{1}{n\,h_1} \sum_{j \in A_i} \left(1 - \frac{t_j - t_i}{h_1}\right)}{1 - \frac{i}{n+1}} \quad . \tag{64}$$

Test Results

The empirical estimator, given in (36), the simple rectangular-window estimator of (54), and the simple triangular-window estimator of (64) were used on a set of randomly generated data from a known truncated exponential distribution and on a set of 200 times-to-death of people from the Los Angeles area. The time-to-death data of people was obtained from the Los Angeles Public Health Cooperative Data File. We wish to thank Dr. Anne Coulson of the School of Public Health, UCLA, for her assistance in obtaining a random sample of 200 people from that file. The date of birth and month and day of death in 1970 was received from cards; if the year of birth was not given, the age at death was subtracted from 1970 and used for the year of birth. The birth date and death date were converted to days and subtracted to give time-to-death in days. A FORTRAN subroutine, CALL DATE (I YEAR, J MONTH, K YEAR, N DAYS), supplied by Steven Chasen of the Health Sciences Computing Facility was used to make the conversion and account for leap years.

The 200 exponential, random numbers were generated by a FORTRAN subroutine, CALL R GAMMA (1.,15000.,200,V), developed by Frane and Murthy

[1973]. The mean was 15,000 days and the variables were truncated at 43,800 days. The hazard is constant at $\lambda = 6.7.10^{-5}$, with a spike at 120 years.

A sort routine, CALL SORT (A,1,200) developed by Steven Chasen was used to put the data in ascending order.

Programs for the estimators were tested for numerical correctness by using uniformly generated data with $t_i = 182(i - 1)$, $1 \leq i \leq 200$; the hazard is $1/(c - t)$, $0 \leq t \leq c$, with $c = 36{,}218$ days and $\lambda(0) = 2.75.10^{-5}$.

The results of the three estimators are shown in Figures 6 - 12. The time-to-death in days was divided by 365 to give time-to-death in years and the hazard was multiplied by 10^5 to provide clarity in the presentation. Various window widths, h, were used; the effect of h on the estimate is shown in Figure 5.

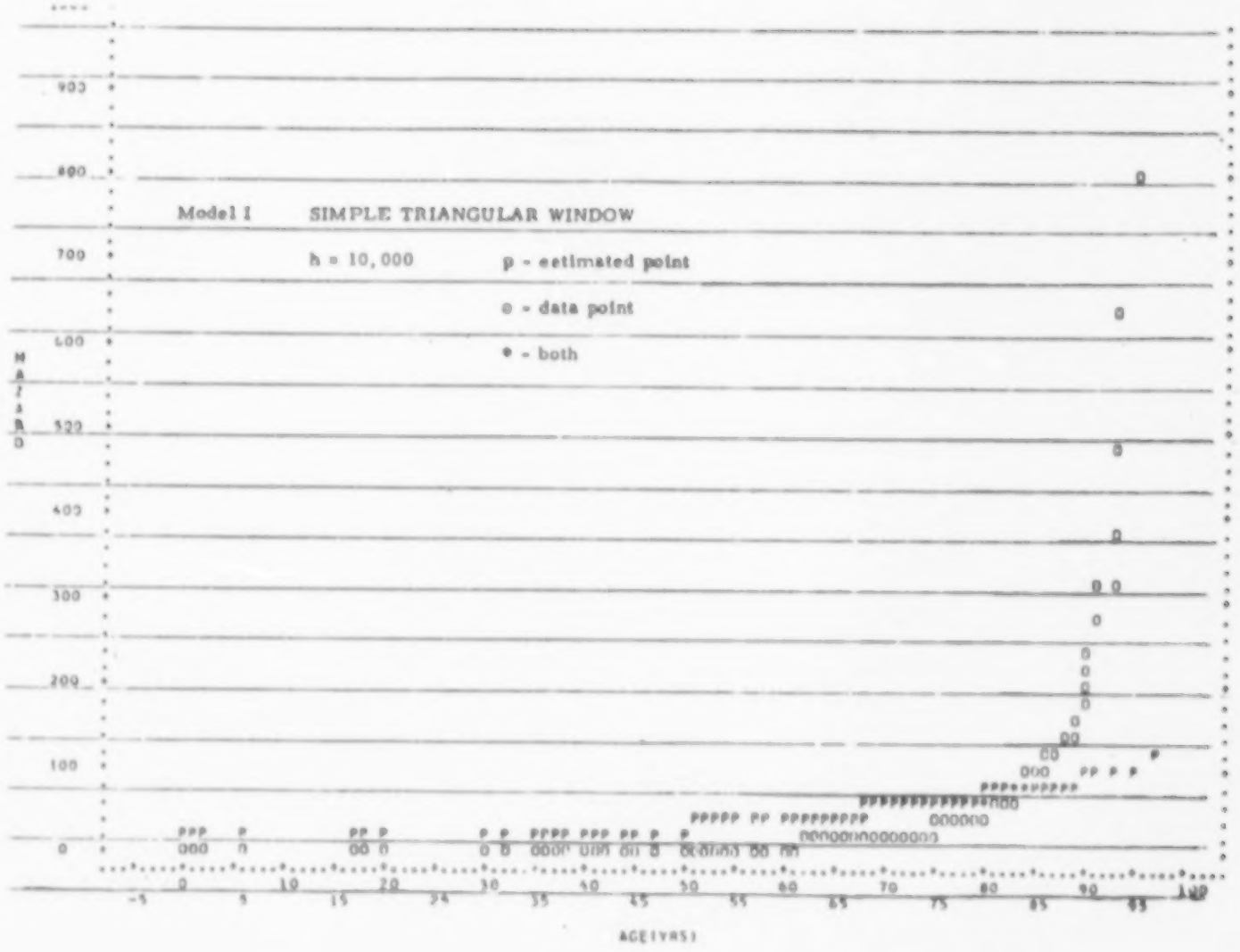

Model I SIMPLE TRIANGULAR WINDOW
h = 10,000
p - estimated point
o - data point
* - both
MAHAD
AGE(YRS)

Figure 2

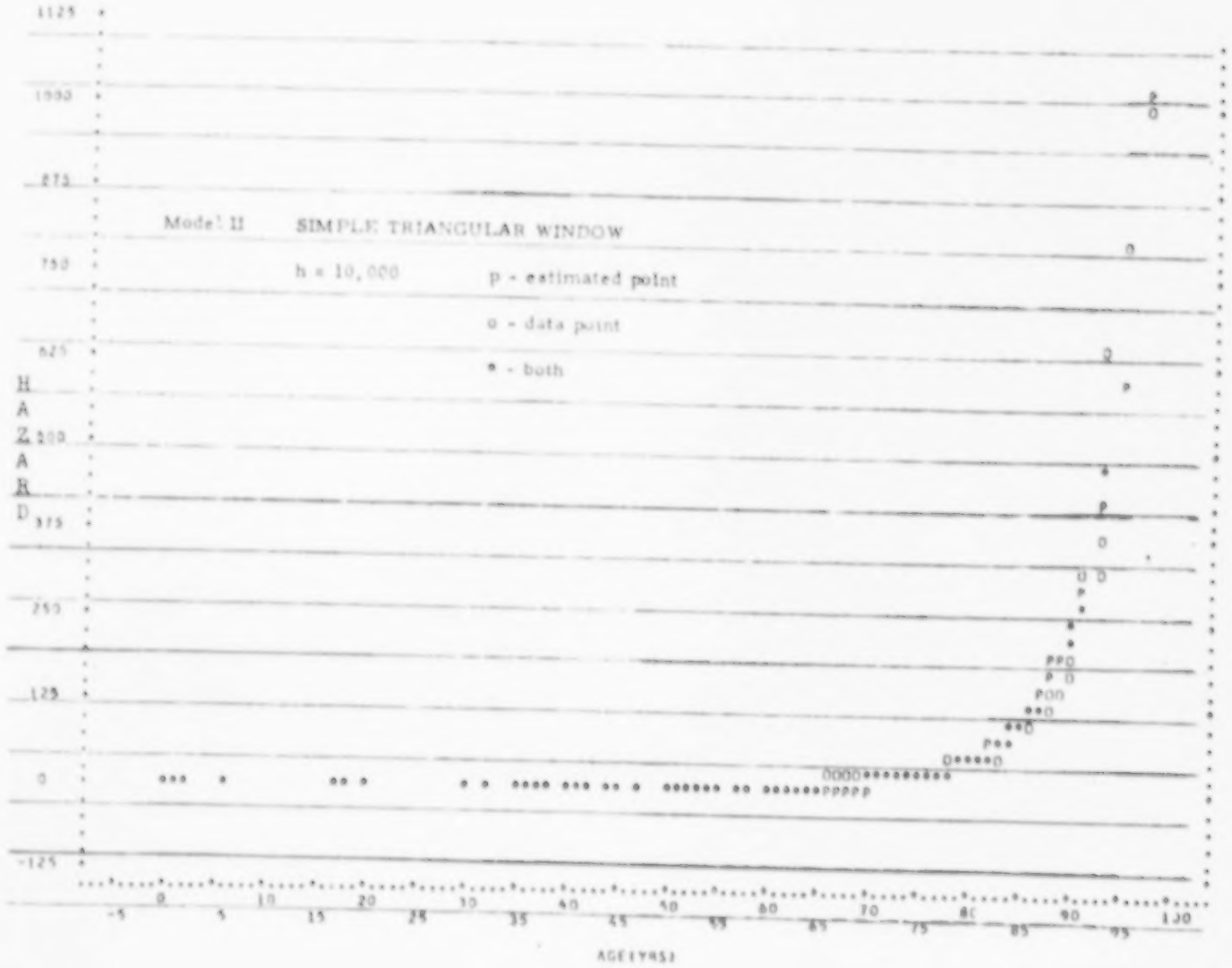

Model II
SIMPLE TRIANGULAR WINDOW
h = 10,000
p - estimated point
o - data point
* - both
HAZARD
AGE(YRS)

Figure 3

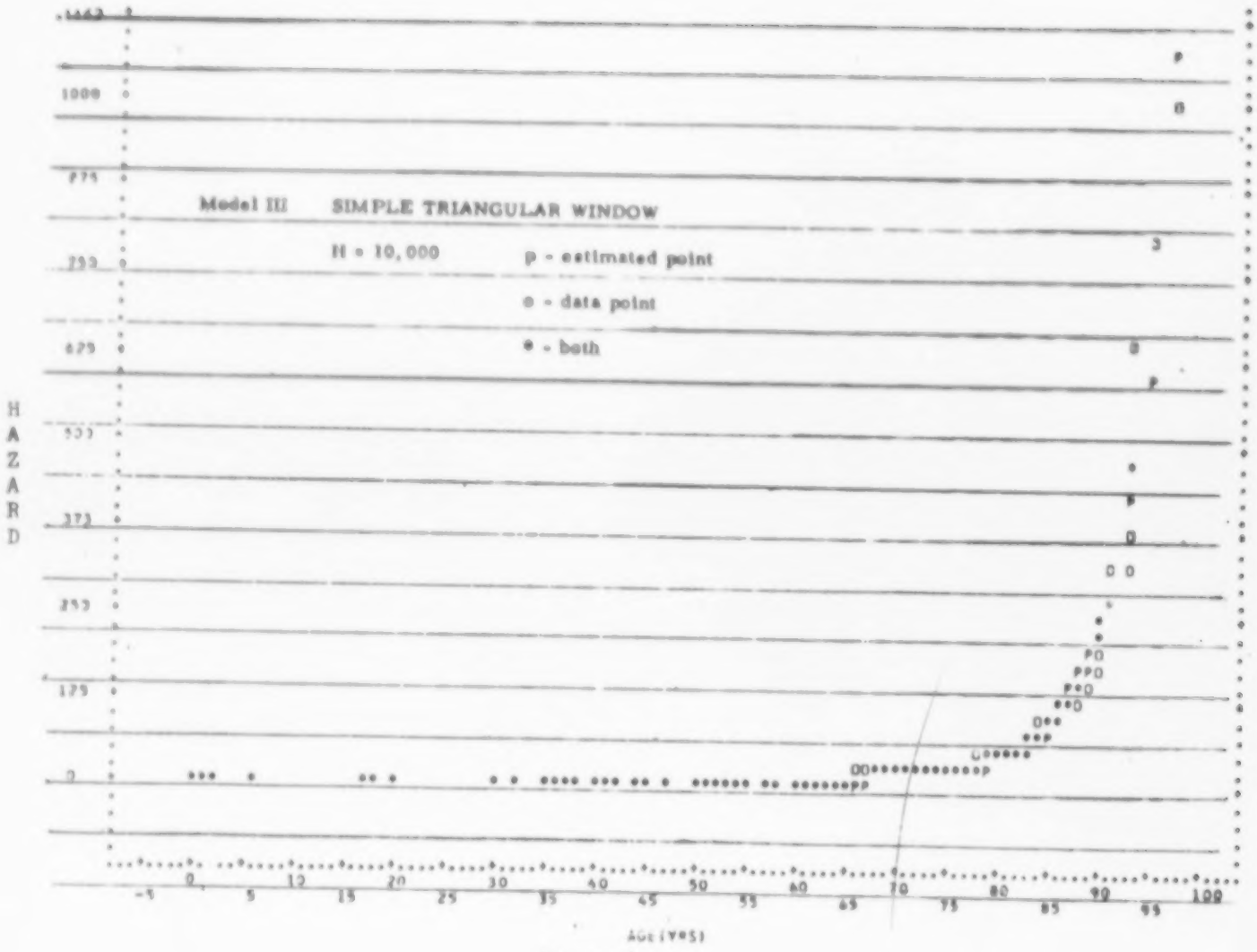
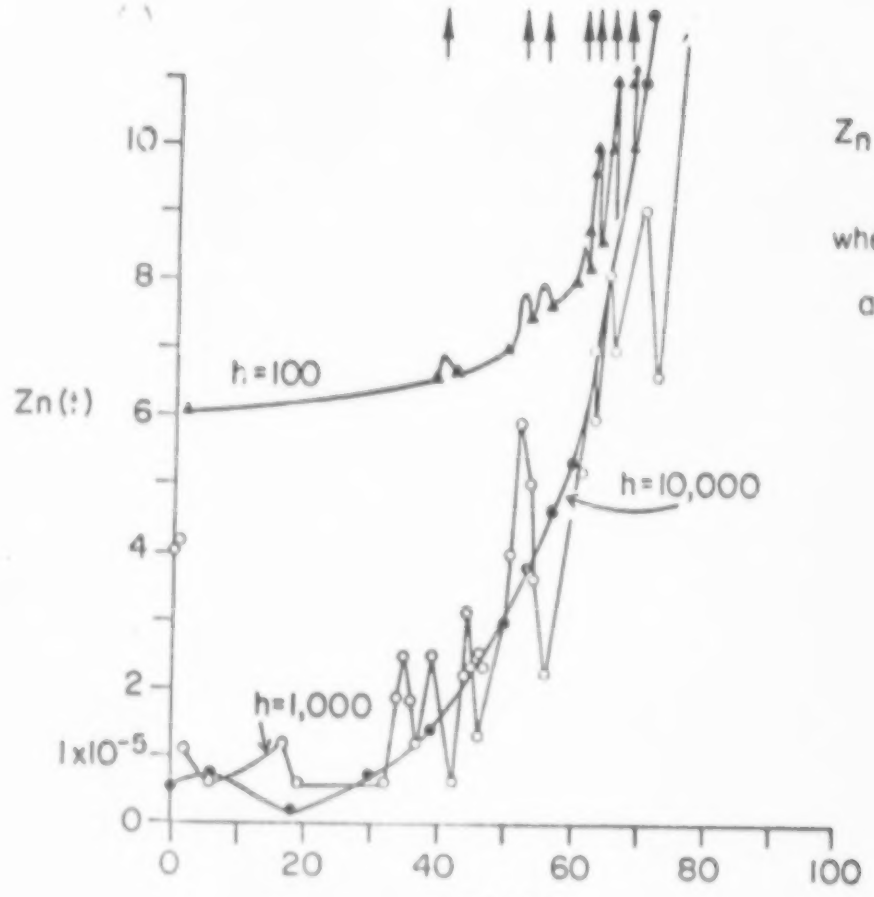

Model III SIMPLE TRIANGULAR WINDOW
N = 10,000
p - estimated point
o - data point
* - both
HAZARD
AGE (YRS)

Figure 4



TIME-TO-DEATH IN LOS ANGELES

$$Z_n(t_i) = \frac{1}{2h,n} \sum_{j \in A_i} 1 \cdot \left(1 - \frac{1}{n+1}\right)$$

where $A_i = \{j: |t_j - t_i| \leq h_i\}$

and $h_i = h/\log_{10} n$

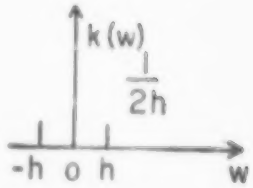


Figure 5

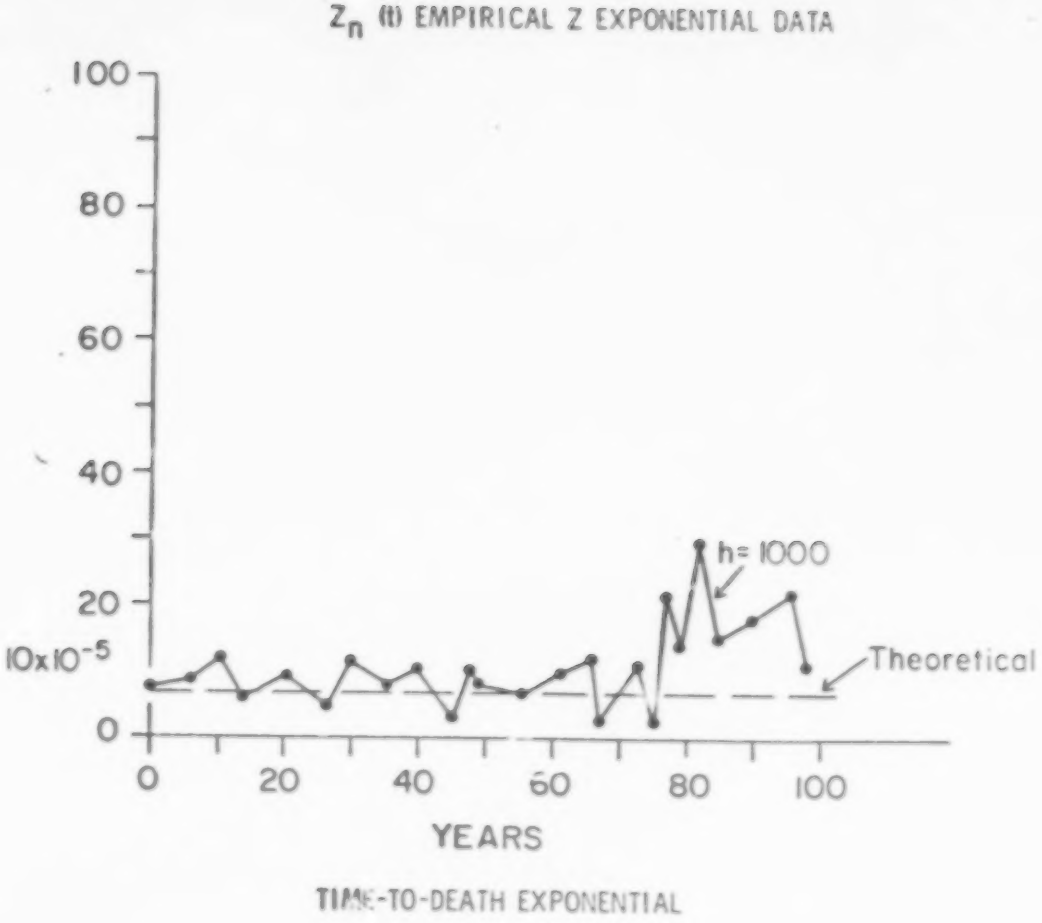

Z_n (t) EMPIRICAL Z EXPONENTIAL DATA
100
80
60
40
20
10x10⁻⁵
0
h= 1000
Theoretical
0
20
40
60
80
100
YEARS
TIME-TO-DEATH EXPONENTIAL

Figure 6

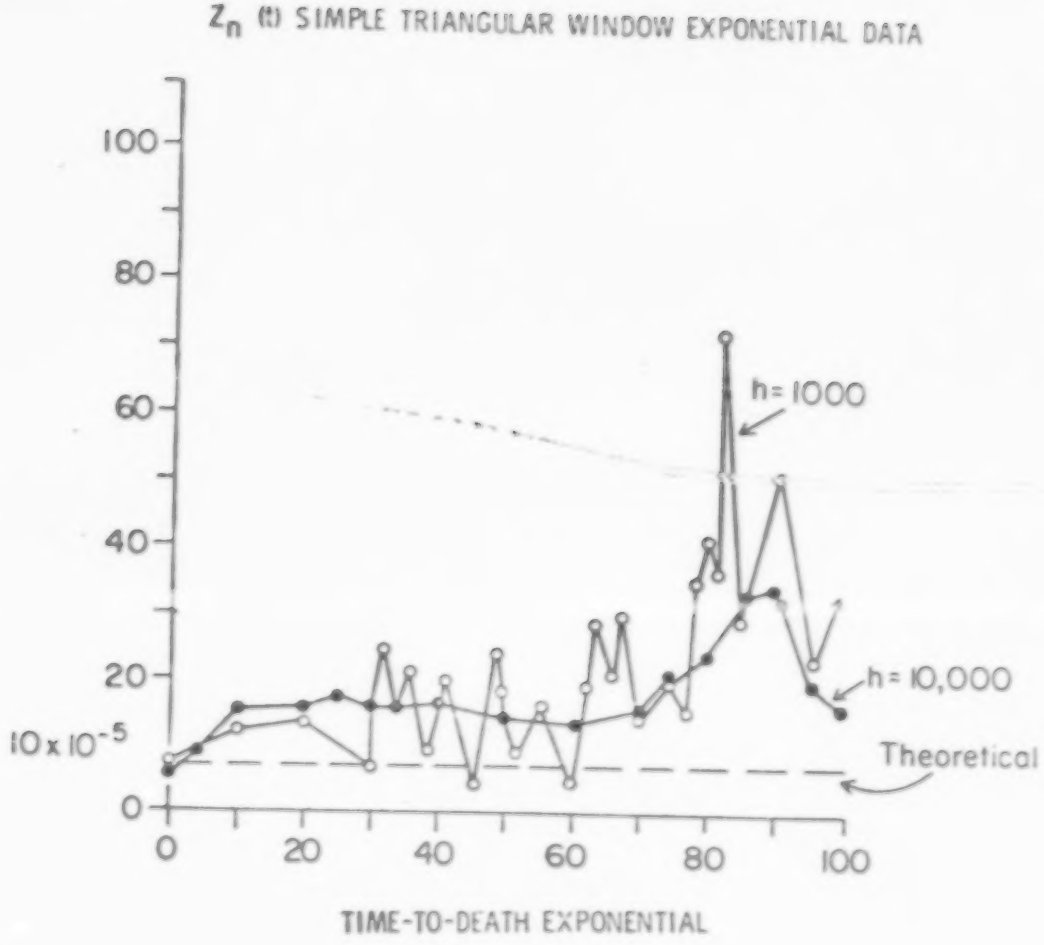

Z_n (t) SIMPLE TRIANGULAR WINDOW EXPONENTIAL DATA
100
80
60
40
20
10×10^{-5}
0
0
20
40
60
80
100
h= 1000
h= 10,000
Theoretical
TIME-TO-DEATH EXPONENTIAL

Figure 7

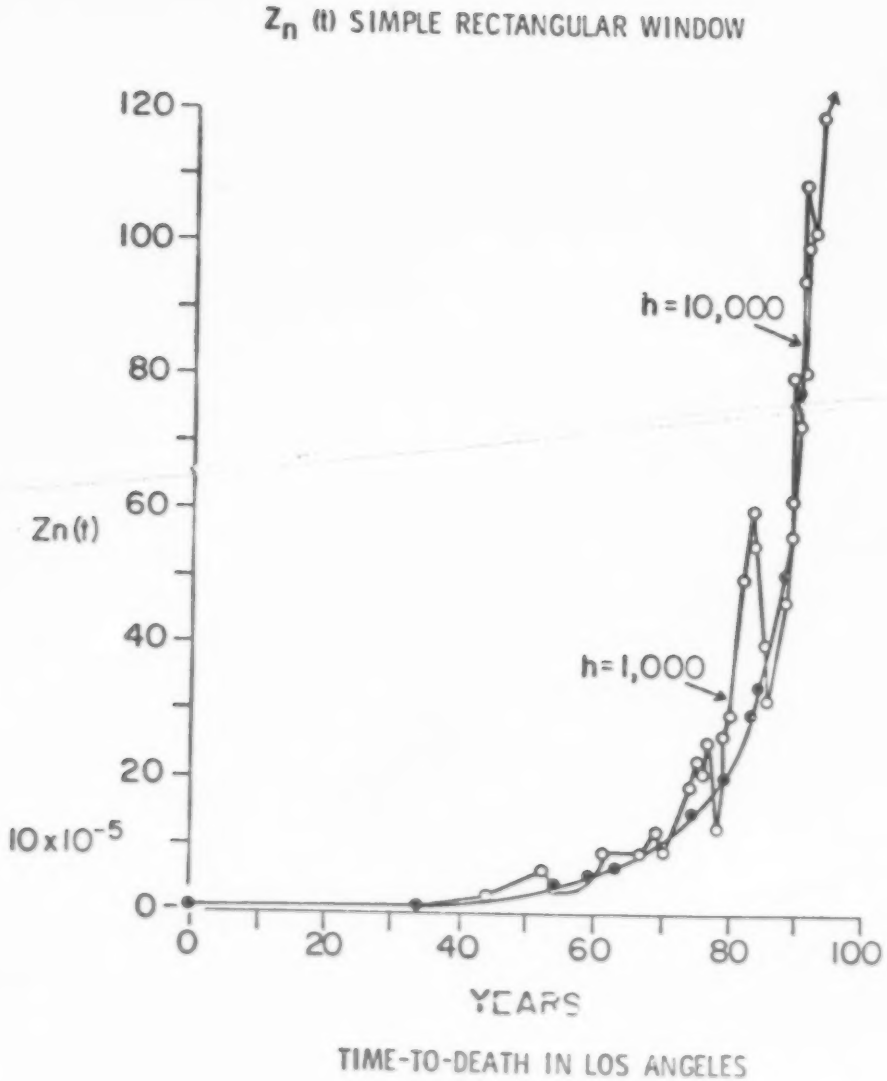

Z_n (t) SIMPLE RECTANGULAR WINDOW
120
100
80
60
40
20
0
Zn(t)
10×10^{-5}
h=10,000
h=1,000
0
20
40
60
80
100
YEARS
TIME-TO-DEATH IN LOS ANGELES

Figure 8

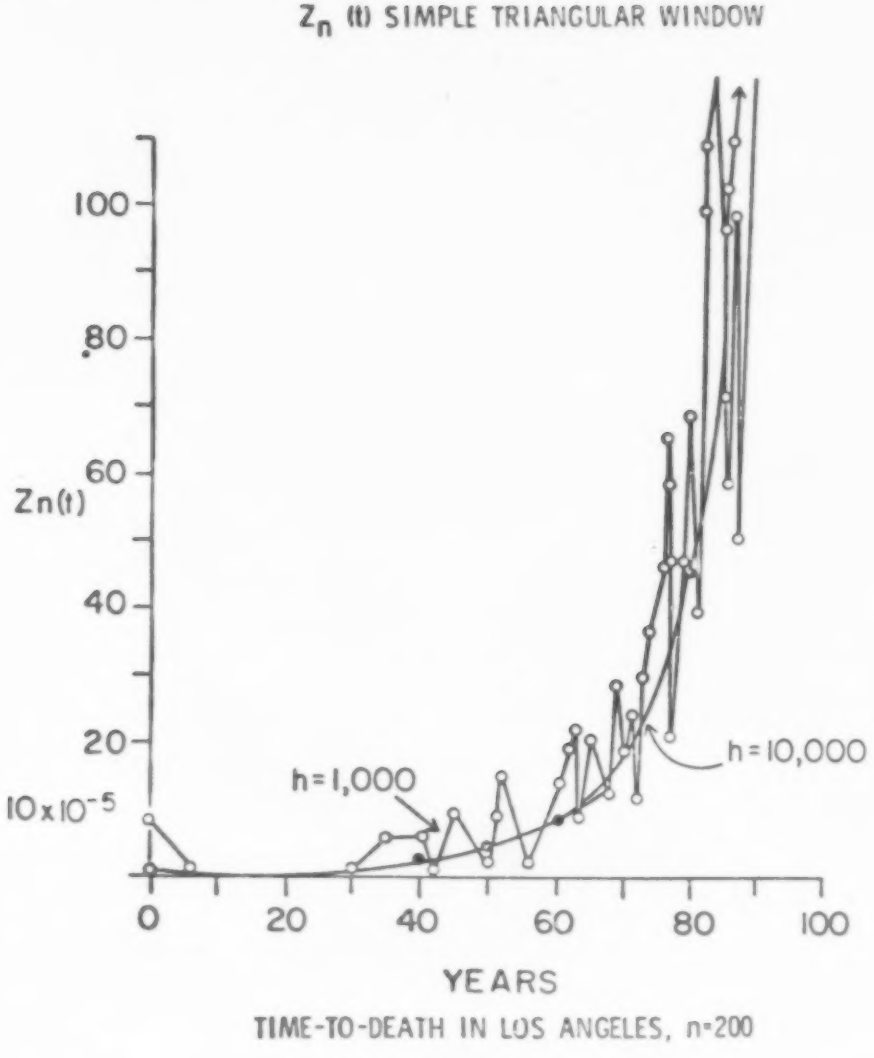

Z_n (t) SIMPLE TRIANGULAR WINDOW
100
80
60
40
20
Zn(t)
10×10^{-5}
h=1,000
h=10,000
0
20
40
60
80
100
YEARS
TIME-TO-DEATH IN LOS ANGELES, n=200

Figure 9

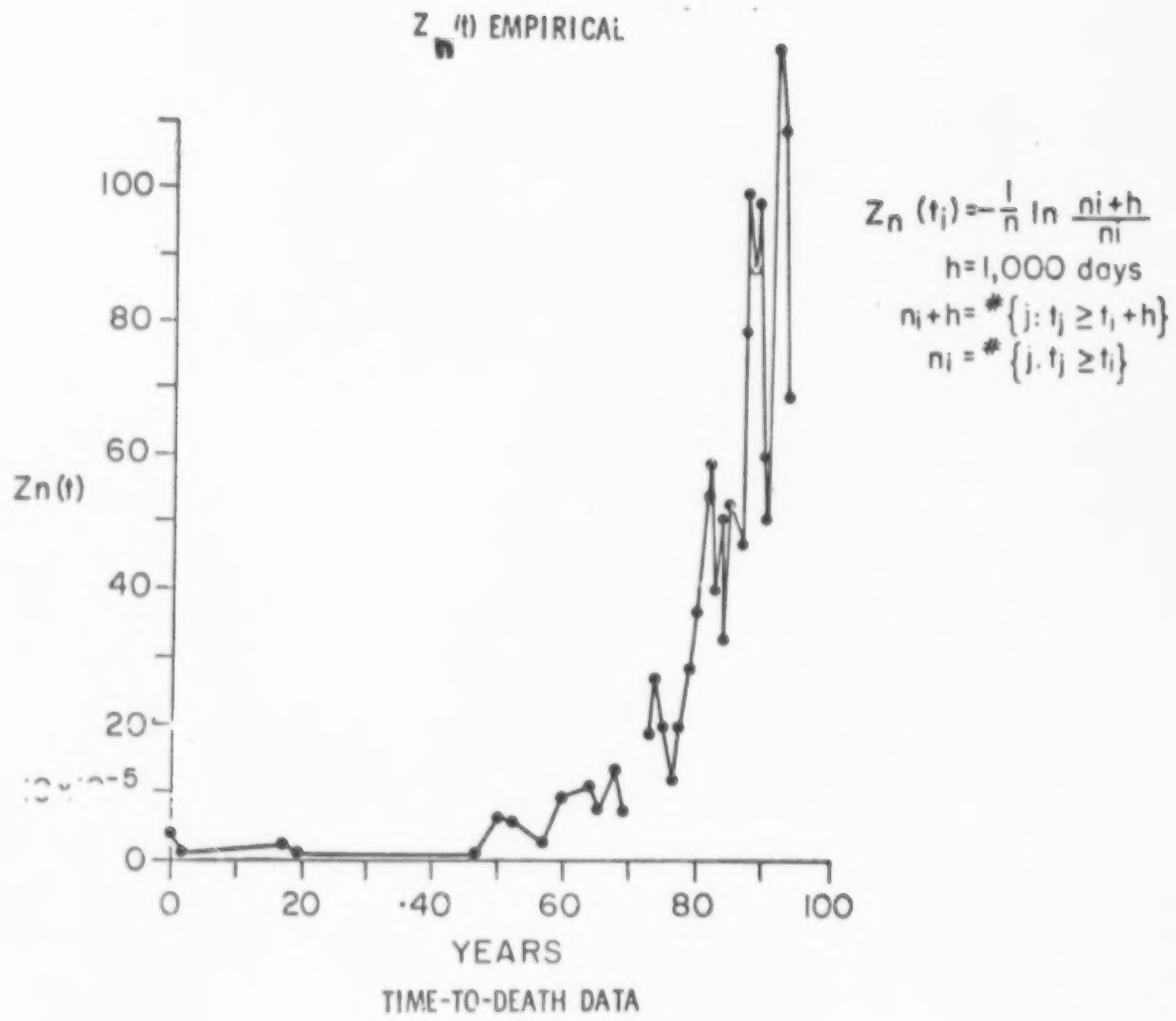
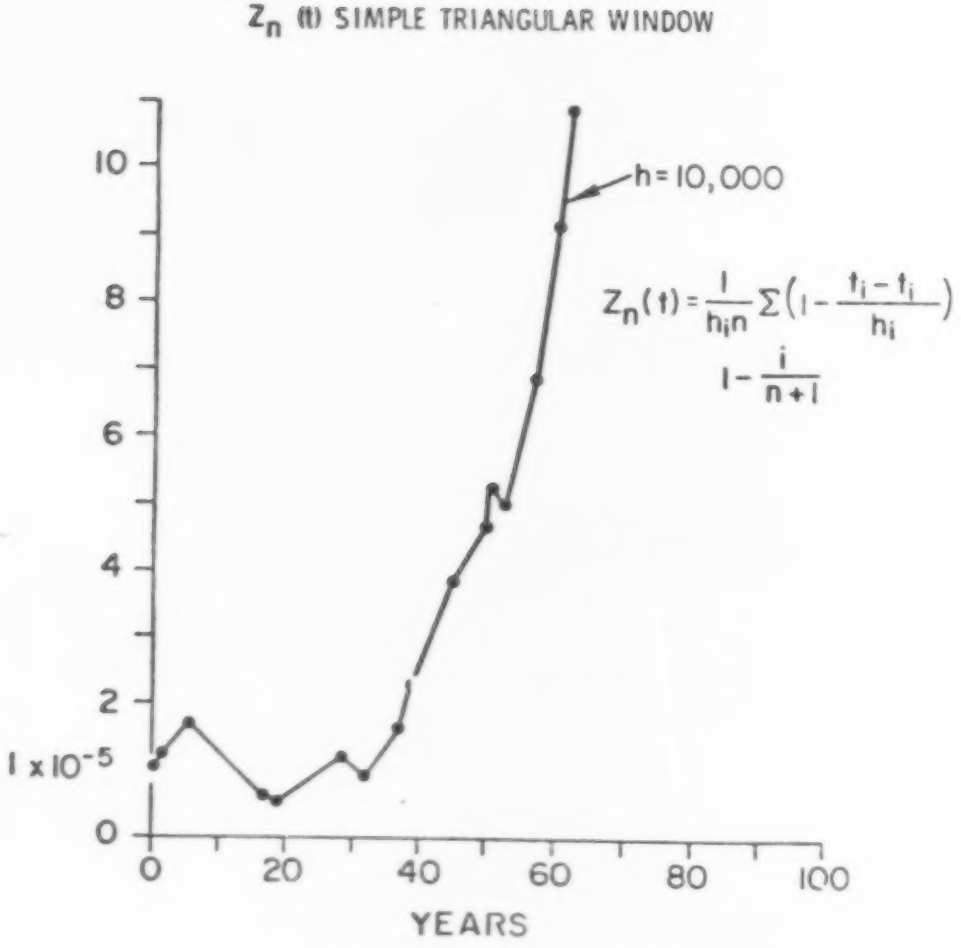

$Z_n(t)$ EMPIRICAL
100
80
60
40
20
0
Zn(t)
0
20
40
60
80
100
YEARS
TIME-TO-DEATH DATA
$Z_n(t_i) = -\frac{1}{n} \ln \frac{n_i + h}{n_i}$
h = 1,000 days
$n_i + h = \#\{j : t_j \geq t_i + h\}$
$n_i = \#\{j . t_j \geq t_i\}$

Figure 10



TIME-TO-DEATH IN LOS ANGELES

Figure 11

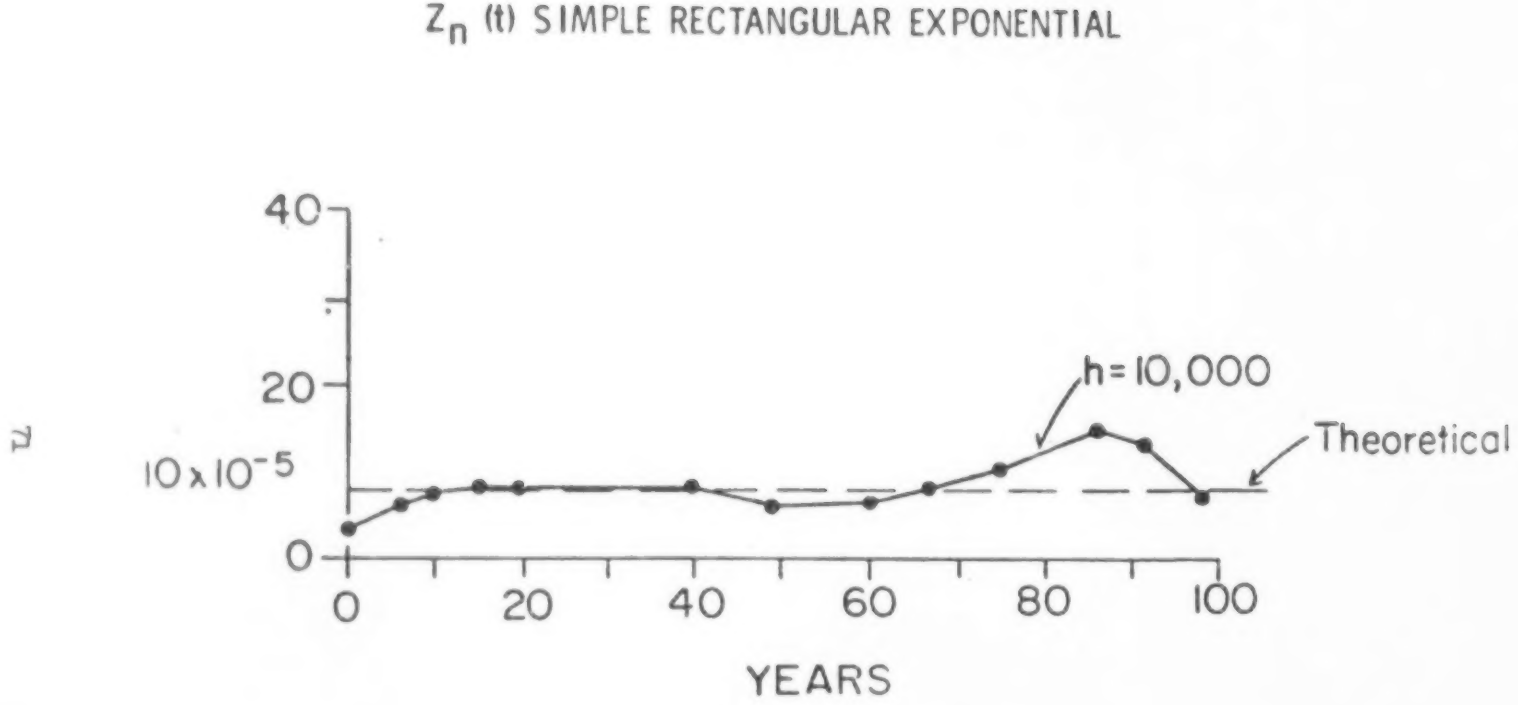

Z_n (t) SIMPLE RECTANGULAR EXPONENTIAL
40
20
10×10^{-5}
0
h = 10,000
Theoretical
0
20
40
60
80
100
YEARS
TIME-TO-DEATH EXPONENTIAL

Figure 12.

CHAPTER 4

Applications of the Models and an Example

Real Life Examples to Which the Model Can Be Applied

The mortality rates corresponding to many real life situations fall within the domain of our models. From the practical point of view, we distinguish among three categories of mortality rates, A, B and C, depending on the shape. Category A is marked by a decreasing risk function, followed by an increasing one. Category B is purely decreasing, while category C is increasing.

In the biological field, the classical example of a bath-tub-hazard curve -- category A -- is the survival data of humans (as well as other animals). It is characterized by: high infant mortality during the first year of life with slowly decreasing risk from age 1 through 4, and the lowest mortality rate starting from age 5 to around 35; increasing risk is seen as age increases above the 35 year level (see Figure 1 of Chapter 3).

This category can be used for survival data pertaining to many varied diseases, as long as they have this general pattern. For example, Gehan's [1969] study was an application of category A, in which $F(t)$ is called the distribution function associated with the mortality rate $\lambda(t)$ and

$$S(t) = 1 - F(t) = e^{y(0)-y(t)} \tag{1}$$

is called the survivorship function. The density function corresponding to (1) is given by

$$f(t) = \frac{dF(t)}{dt} \quad . \tag{2}$$

It is clear from Equations (1) and (2) that there is a one-to-one correspondence between the mortality rate on the one hand and the distribution and survivorship functions on the other. If we can discover mortality functions $\lambda(t)$ with the bath tub property and obtain the integral $y(t) = \int_0^t \lambda(t)dt$ in a closed form, Equations (1) and (2) will give the corresponding distribution and survivorship functions, respectively, in closed form. In what follows, we obtain families of distributions whose corresponding mortality rates have the above desired properties.

Statistical Models with Bath-Tub Property

MODEL I: This class of distributions is characterized by the hazard rate

$$\lambda(t) = \frac{a}{1 + bt} + c\, dt^{d-1}, \qquad t \geq 0 \tag{3}$$

where the parameters $a,b,c \geq 0$, $d \geq 1$. Some special cases of this model are: when $c = 0$,

$$\lambda(t) = \frac{a}{1 + bt} \quad , \tag{4}$$

a decreasing function of time.

When

$$\left.\begin{aligned} a = 0,\ d = 1, \quad & \lambda(t) = c \\ c = 0,\ b = 0, \quad & \lambda(t) = a \end{aligned}\right\} \qquad (5)$$

Here, the mortality rate reduces to a constant and we have an exponential distribution. When $b = 0$, $d = 2$, we have the form described by Kodlin [1967]: that form the observed data on survival of males with angina pectoris, the death rate per yeat is highest in the first year after diagnosis. For the next nine years, the hazard function remains fairly constant and increases somewhat later.

Category B curves characterize situations where there is an initial high risk which decreases slowly as time progresses. Survival data on patients after surgery, transplant operation, or other medical treatments that would improve the chances of survival, belong to this category.

Category C is marked by an increasing risk function. Experiments that introduce harm to initially healthy animals, or sick control groups that are unattended are good examples for this. Garg, et al [1970] and Kodlin [1967] gave numerical examples of this type in their papers.

Graphics Program

We shall now demonstrate the versatility of the models considered in this chapter, with respect to their ability to describe the various mortality rates that one encounters in real life. To this end, an interactive graphics program was written in such a way that a person at the console can select one of the models at a time, and type in different values of the parameters

through a keyboard. The corresponding hazard function will then be displayed on the screen. If desired, different increments of the parameters can also be typed in and plots of these curves will be displayed one after the other, as the parameters increase or decrease by the specified amount each time. Thus, the characteristics of each model can be fully illustrated easily.

Maximum Likelihood Procedure for Estimation

The method of maximum likelihood is discussed here. Since similar techniques are applied to all three models, more detailed explanations will be given for Model I, while only brief summaries of equations will be given for Models II and III.

MODEL I

Let $t_1, t_2, \ldots, t_n$ be a random sample of size N from the distribution $F(t)$ given by (4). The likelihood of the observed sample is

$$L = \prod_{i=1}^{N} \frac{1}{(1+bti)^{a/b}} e^{-cti^d} \left[cdti^{d-1} + \frac{a}{1+bti} \right] \tag{6}$$

$$\ln L = -\frac{a}{b} \sum_{i=1}^{N} \ln(1+bti) - c \sum_{i=1}^{n} ti^d + \sum_{1}^{N} \ln\left[cdti^{d-1} + \frac{a}{1+bti} \right] . \tag{7}$$

The following equations can be solved, using $\hat{a},\hat{b},\hat{c}$ and $\hat{d}$ that will maximise (7); the second partial derivatives given by:

$$\frac{\partial^2 \ln L}{\partial a^2} = -b^2 \sum_1^N \frac{t_i^{2(b-1)}}{[\lambda(t_i)]^2} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial a \partial b} = -\sum_1^N t_i^b \ln t_i + \sum_1^N \frac{t_i^{b-1}(1 + b \ln t_i)}{\lambda(t_i)} - ab \sum_1^N \frac{t_i^{2(b-1)}}{[\lambda(t_i)]^2} \left[1 + b \ln t_i\right] \quad ;$$

$$\frac{\partial^2 \ln L}{\partial a \partial c} = bd \sum_1^N \frac{t_i^b}{[\lambda(t_i)]^2 (1 + ct_i)^2} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial a \partial d} = -b \sum_1^N \frac{t_i^{(b-1)}}{[\lambda(t_i)]^2 (1 + ct_i)} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial b^2} = -a \sum_1^N t_i^b (\ln t_i)^2 + a \sum_1^N \frac{t_i^{b-1} \ln t_i (2 + b \ln t_i)}{\lambda(t_i)} - a^2 \sum_1^N \frac{t_i^{2(b-1)} (1 + b \ln t_i)^2}{[\lambda(t_i)]^2} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial b \partial c} = ad \sum_1^N \frac{t_i^b (1 + b \ln t_i)}{[\lambda(t_i)]^2 (1 + ct_i)^2} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial b\, \partial d} = -a \sum_{1}^{N} \frac{t_i^{b-1}(1 + b \ln t_i)}{[\lambda(t_i)]^2 (1 + ct_i)}$$

$$\frac{\partial^2 \ln L}{\partial c^2} = \frac{2d}{c} \sum_{1}^{N} \frac{t_i}{1 + ct_i} - \frac{2d}{c^3} \sum_{1}^{N} \ln(1 + ct_i) + \frac{d}{c} \sum_{1}^{N} \frac{t_i^2}{(1 + ct_i)^2}$$

$$+ 2d \sum_{1}^{N} \frac{t_i^2}{\lambda(t_i)(1 + ct_i)^3} - d^2 \sum_{1}^{N} \frac{t_i^2}{[\lambda(t_i)]^2 (1 + ct_i)^4} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial c\, \partial d} = \frac{1}{c^2} \sum_{1}^{N} \ln(1 + ct_i) - \frac{1}{c} \sum_{1}^{N} \frac{t_i}{1 + ct_i} - \sum_{1}^{N} \frac{t_i}{\lambda(t_i)(1 + ct_i)^2}$$

$$+ d \sum_{1}^{N} \frac{t_i}{[\lambda(t_i)]^2 (1 + ct_i)^3} \quad ;$$

$$\frac{\partial^2 \ln L}{\partial d^2} = -\sum_{1}^{N} \frac{1}{[\lambda(t_i)]^2 (1 + ct_i)^2} \quad ;$$

$$\frac{\partial \ln L}{\partial a} = -\frac{1}{b} \sum_{1}^{N} \ln(1 + ct_i) + \sum_{1}^{N} \frac{1}{\lambda(t_i)(1 + bt_i)} = 0 \quad ;$$

$$\frac{\partial \ln L}{\partial b} = \frac{a}{b^2} \sum_{1}^{N} \ln(1 + bt_i) - \frac{a}{b} \sum_{1}^{N} \frac{t_i}{1 + bt_i} - a \sum_{1}^{N} \frac{t_i}{\lambda(t_i)(1 + bt_i)} = 0 \quad ;$$

$$\frac{\partial \ln L}{\partial c} = -\sum_{1}^{N} t_i^d + d \sum_{1}^{N} \frac{t_i^{d-1}}{\lambda(t_i)} = 0 \quad ;$$

and

$$\frac{\partial \ln L}{\partial d} = -c \sum_{R} t_i^d \ln t_i + c \sum_{1} \frac{t_i^{d-1}(1 + d \ln t_i)}{\lambda(t_i)}$$

where

$$\lambda(t_i) = cdt_i^{d-1} + \frac{a}{1 + bt_i} \quad . \tag{8}$$

With the aid of a computer, these equations can be solved iteratively. Initial estimates can be obtained from nonlinear least-squares fit to the empirical hazard function.

Estimates of the asymptotic variances and covariances of the maximum likelihood estimates may be obtained from the matrix:

$$M = \left[\begin{array}{cccc} -\frac{\partial^2 \ln L}{\partial a^2} & -\frac{\partial^2 \ln L}{\partial a \partial b} & -\frac{\partial^2 \ln L}{\partial a \partial c} & -\frac{\partial^2 \ln L}{\partial a \partial d} \\ -\frac{\partial^2 \ln L}{\partial a \partial b} & -\frac{\partial^2 \ln L}{\partial b^2} & -\frac{\partial^2 \ln L}{\partial b \partial c} & -\frac{\partial^2 \ln L}{\partial b \partial d} \\ -\frac{\partial^2 \ln L}{\partial a \partial c} & -\frac{\partial^2 \ln L}{\partial b \partial c} & -\frac{\partial^2 \ln L}{\partial c^2} & -\frac{\partial^2 \ln L}{\partial c \partial d} \\ -\frac{\partial^2 \ln L}{\partial a \partial d} & -\frac{\partial^2 \ln L}{\partial b \partial d} & -\frac{\partial^2 \ln L}{\partial c \partial d} & -\frac{\partial^2 \ln L}{\partial d^2} \end{array} \right]^{-1} \tag{9}$$

evaluated at $a = a$, $b = b$, $c = c$ and $d = d$.

MODEL II

The equations involved in solving for Model II are, briefly:

$$\ln L = \sum_{1}^{N} \ln(ab\, e^{-bt_i} + cdt_i^{d-1}) - \sum_{1}^{N} a(1 - e^{-bt_i}) - \sum_{1}^{N} ct_i^d \qquad (10)$$

with

$$\frac{\partial \ln L}{\partial a} = b \sum_{1}^{N} \frac{e^{-bt_i}}{ab\, e^{-bt_i} + cdt_i^{d-1}} - \sum_{1}^{N} (1 - e^{-bt_i}) = 0 \quad ;$$

$$\frac{\partial \ln L}{\partial b} = a \sum_{1}^{N} \frac{e^{-bt_i}(1 - bt_i)}{ab\, e^{-bt_i} + cdt_i^{d-1}} - a \sum_{1}^{N} t_i\, e^{-bt_i} = 0 \quad ;$$

$$\frac{\partial \ln L}{\partial c} = d \sum_{1}^{N} \frac{t_i^{d-1}}{ab\, e^{-bt_i} + cdt_i^{d-1}} - \sum_{1}^{N} t_i^d = 0 \quad ;$$

and

$$\frac{\partial \ln L}{\partial d} = c \sum_{1}^{N} \frac{t_i^{d-1}(1 + d \ln t_i)}{ab\, e^{-bt_i} + cdt_i^{d-1}} - c \sum_{1}^{N} t_i^d \ln t_i = 0 \quad .$$

MODEL III

$$\ln L = \sum_{1}^{N} \ln\left(\frac{a}{1 + bt_i} + cd\, e^{dt_i} \right) - \frac{a}{b} \sum_{k}^{N} \ln\left(1 + bt_i \right) - c \sum_{1}^{N} \left(e^{dt_i} - 1 \right) \qquad (11)$$

with

$$\frac{\partial \ln L}{\partial a} = \sum_{1}^{N} \frac{1}{\left(\frac{a}{1 + bt_i} + cd\, e^{dt_i} \right)\left(1 + bt_i \right)} - \sum_{1}^{N} \frac{\ln(1 + bt_i)}{b} = 0 \quad ;$$

$$\frac{\partial \ln L}{\partial b} = - \sum_{1}^{N} \frac{a t_i}{\left(\frac{a}{1 + bt_i} + cd\, e^{dt_i} \right)\left(1 + bt_i \right)^2} + \frac{a}{b^2} \sum_{1}^{N} \ln\left(1 + bt_i \right) - \frac{a}{b} \sum_{1}^{N} \frac{t_i}{1 + bt_i} = 0 \quad ;$$

$$\frac{\partial \ln L}{\partial c} = \sum_{1}^{N} \frac{d\, e^{dt_i}}{\frac{a}{1 + bt_i} + cd\, e^{dt_i}} - \sum_{k}^{N} \left(e^{dt_i} - 1 \right) = 0 \quad ;$$

and

$$\frac{\partial \ln L}{\partial d} = \sum_{1}^{N} \frac{c\, e^{dt_i}(1 + dt_i)}{\frac{a}{1 + bt_i} + c\, de^{dt_i}} - c \sum_{1}^{N} t_i\, e^{dt_i} = 0 \quad .$$

A Numerical Example

A hypothetical cohort of 1,000 people born in the United States in the year 1900 was made up, using information recorded in Vital Statistics of the United States. (The probability of dying for ages ≥ 70 was approximated using the values in the latest 1967 of the aforementioned reference.) The survival data are presented below and the empirical hazard function was calculated using

$$\lambda(t_{mi}) = \frac{2q_i}{h_i(1 + p_i)}$$

where

t_{mi} = midpoint of the i^{th} interval;

h_i = width of the i^{th} interval;

q_i = conditional proportion dying in the i^{th} interval; and

p_i = conditional proportion surviving in the i^{th} interval;

as described by Kimball [1960].

Nonlinear least-squares fits to the empirical hazard function were attempted for all three models. (Computations were performed using BMDX85, Nonlinear Least Squares, a package program written at the Health Sciences Computing Facility, UCLA.) The estimated parameters and hazard functions are shown in Table 2.

It was found that Model I seems to fit the data best, so the method

of maximum likelihood was performed, giving estimates as follows:

$$\hat{a} = 1.5942 \text{ E } 01$$
$$\hat{b} = 2.612 \text{ E } 02$$
$$\hat{c} = 8.635 \text{ E-}11 \qquad (12)$$
$$\hat{d} = 6.299 \text{ E } 00$$

The value of the log likelihood (Eq. 7), using maximum likelihood estimates is -0.4378E04, compared to -0.4413E04 using the least-squares estimates.

TABLE 1

U.S. SURVIVAL DATA

*$\lambda(t) = 2q/(h(1+p))$ where h = year interval, according to Kimball's method.

Age	N	Dead	q	p	pp	$\lambda(t)^*$
0- 0.99	1000	162	0.1620	0.8380	0.83800	0.176270
1- 4.99	838	39	0.0465	0.9535	0.79903	0.011900
5-14.99	799	23	0.0288	0.9712	0.77602	0.002922
15-24.99	776	38	0.0490	0.0951	0.73800	0.005023
25-34.99	738	35	0.0474	0.9526	0.70302	0.004855
35-44.99	703	37	0.0526	0.9474	0.66604	0.005402
45-54.99	666	57	0.0856	0.9144	0.60903	0.008942
55-64.99	609	106	0.1740	0.8260	0.50306	0.019058
65-74.99	503	189	0.3757	0.6243	0.31406	0.046259
75-84.99	314	248	0.7898	0.2102	0.06602	0.130534
85-94.99	66	66	1.0000	0.0000	0.00000	0.200000

TABLE 2

U.S. SURVIVAL DATA

(Hazard Function)*

Midpoint Year	Observed	Model I	Model II	Model III
0.5	0.17630	0.17307	0.17630	0.17453
3.0	0.01190	0.02906	0.01191	0.02922
10.0	0.00292	0.00873	0.00001	0.00877
20.0	0.00502	0.00445	0.00010	0.00439
30.0	0.00485	0.00360	0.00077	0.00292
40.0	0.00540	0.00524	0.00333	0.00219
50.0	0.00894	0.01147	0.01036	0.00175
60.0	0.01906	0.02649	0.02617	0.00146
70.0	0.04626	0.05693	0.05730	0.00126
80.0	0.13052	0.11237	0.11299	0.00244
90.0	0.20000	0.20591	0.20563	0.20087

Model I	Model II	Model III
a = 1.9872E 01	a = 2.8035E-01	a = 3.2761E 01
b = 2.2764E 02	b = 1.0781E 00	b = 3.7340E 02
c = 2.4398E-12	c = 3.9204E-12	c = 1.1531E-20
d = 6.1851E 00	d = 6.0841E 00	d = 4.9992E-01

*Least-squares estimation on $\lambda(t)$.

CHAPTER 5

Computer Program for Graphics

The computer program described in this section was used to display, on a single plot, both the mortality rate and the mean residual lifetime for three realistic models of the mortality rate. This program was written in FORTRAN IV and was run on an IBM 360/91 at the Health Sciences Computing Facility, UCLA.

The mean residual lifetime, $v(t)$, may be defined as

$$v(t) = \int_t^\infty \frac{R(x)}{R(t)}\,dx$$

where $R(t) = 1 - F(t)$ is the survivorship function and $F(t)$ the cumulative distribution function. For Model I, the mortality rate, $\lambda(t)$ and the survivorship function $R(t)$ are

$$\lambda(t) = \frac{a}{1+bt} + c\,dt^{d-1}, \qquad t \geq 0,\ a,b,c \geq 0,\ d \geq 1$$

and

$$R(t) = \frac{1}{(1+bt)^{a/b}}\, e^{-ct^d} .$$

For Model II, the equations are

$$\lambda(t) = abe^{-bt} + cdt^{d-1}, \qquad t \geq 0,\ d \geq 1,\ a,b,c \geq 0$$

and

$$R(t) = e^{-a(1-e^{-bt})-ct^d} .$$

Finally, for Model III the equations are

$$\lambda(t) = \frac{a}{1 + bt} + cde^{dt}, \quad t \geq 0, \quad a,b,c,d \geq 0$$

and

$$R(t) = \frac{1}{(1 + bt)^{a/b}} e^{-c(e^{dt}-1)} .$$

The mortality functions for the three models are computed in the subroutine HAZ listed below. The value of LM may be 1, 2 or 3, and refers to the appropriate model number with HAZ being the mortality rate value at a given time.

```
      FUNCTION HAZ(T)
      DOUBLE PRECISION A,B,C,D,Z,X
      COMMON LM
      COMMON/PARS/A,B,C,D
      X=DBLE(T)
      GO TO (1,2,3),LM
    1 CONTINUE
      Z=A/(1.+B*X)
      HAZ=Z+C*D*X**(D-1.)
      RETURN
    2 CONTINUE
      Z=A*B*DEXP(-B*X)
      HAZ=Z+C*D*X**(D-1.)
      RETURN
    3 CONTINUE
      Z=A/(1.+B*X)
      HAZ=Z+C*D*DEXP(D*X)
      RETURN
      END
```

The survivorship functions for the three models are computed in the double precision subroutine FUN listed below. The value of LM may be 1,2 or 3 and refers to the appropriate model number with FUN being the survivorship

function value at a given time.

```
      DOUBLE PRECISION FUNCTION FUN(X)
      DOUBLE PRECISION X,A,B,C,D,EX,EX
      COMMON LM
      COMMON/PARS/A,B,C,D
      GO TO (1,2,3),LM
    1 CONTINUE
      EX=C*X**D
      IF(EX.GT.100.) EX=100.
      EX=DEXP(-EX)/(1.0D0+B*X)**(A/B)
      FUN=EX
      RETURN
    2 CONTINUE
      EX=B*X
      IF(EX.GT.170.) EX=170.
      EX=A (1.0D0-DEXP(-EX))+C*X**D
      IF(EX.GT.100.) EX=100.
      FUN=DEXP(-EX)
      RETURN
    3 CONTINUE
      EX=D*X
      IF(EX.GT.100.) EX=100.
      EX=C (DEXP(EX)-1.0D0))
      IF(EX.GT.100.) EX=100.
      EX=DEXP(-EX)
      FUN=EX/(1.0D0+B*X)**(A/B)
      RETURN
      END
```

Within the main program, the appropriate model number is read, the values for the parameters a,b,c and d of the model are read and values for the mean residual lifetime are computed.

Finally, the specified information is rpinted along with the computed time values, mean residual lifetime values, v, and the mortality rate values Z; the main program calls a subroutine GPLOT which plots Z and v at various times. The routines used in GPLOT were developed at the Health

Sciences Computing Facility, U.C.L.A.

Sample output from this program is included after the listings of the main program and subroutine GPLOT.

```
      DIMENSION V(400), TIME(400), G(400)
      DIMENSION LM(3)
      DOUBLE PRECISION DA,DB,DL,DI,DFLA
      DIMENSION Z(400), X(400),Y(400)
      DOUBLE PRECISION  A, B, C, D
      DOUBLE P ECISION DTES
      COMMON /PARMS/ A,B,C,D
      DATA CH/'V','Z','*'/
      COMMON L1
10    READ(5,100)  LM
      IF(LM .LT. 0) GO TO 99
17    READ(5,111) A,B,C,D
111   FORMAT(4F10.5)
      WRITE(6,120) LM,A,B,C,D
      IF(A .EQ. -1.0) GO TO 10
      IMIN=0.
      IMAX=40.
      STEP=2.
      AREA=0.000
      XSTEP=100
      XNUM=(IMAX-IMIN)/STEP
      IU=IMAX
      DO 11 I=1,1000
      I=DBLE(IU)
      IF(FUN(I) .LE. 1.00-6) GO TO 20
21    IU=IU+STEP
      WRITE(6,112) IU
112   FORMAT(' FAIL TOO BIG AT IU=',F10.5)
20    CONTINUE
      XX=STEP/.STEP
      DA=DBLE(AX)
      H=FUN(I)
      NUM=0
      IEND=1.0000
      DO 3 I=1,1000
      TEST=IMAX-STEP*NUM
      DTES=DMIN1(TEST)
      IF(DABS(I-DTES) .GT. 1.0D-2) GO TO 4
      J=XNUM+1 -NUM
      IF(H .LE. 1.0E-6) GO TO 12
      V(J)=AREA/H
```

```
      GO TO 13
   12 V(J)=0.
   13 CONTINUE
      Z(J)=FAZ(TEST)*1.0E2
      TIME(J)=TEST
      NUM=NUM+1
      IF(NUM .GT. NNUM) GO TO 69
    2 T=T-DX
      L=F(T)
      E=(L+R)/2.
      R=L
    1 AREA=AREA+E*DX


      WRITE(6,121) NUM
  121 FORMAT('  NO CONV AT NUM', I10)
   69 CONTINUE
      WRITE(6,175)
  175 FORMAT('0   I',10X,'TIME',10X,'V',15X,'Z')
      DO 22 I=1,NUM
      Y(I)=V(I)
      X(I)=TIME(I)
      G(I)=1.
      IT2=I+NUM
      Y(IT2)=Z(I)
      X(IT2)=TIME(I)
      G(IT2)=2.
      WRITE(6,140) I,X(I),V(I),Z(I)
  140 FORMAT(1H , I5, 3F15.5)
   22 CONTINUE
      WRITE(6,150)
  150 FORMAT('1  Z,V  VERSUS  T')
      NK=NUM*2
      CALL CPLOT(X,Y,G,NK,CH,2)
      GO TO 17
  100 FORMAT(I5)
  120 FORMAT('1  MODEL', I5, '  A,  B,  C,  E=', 4F10.5)
   90 CONTINUE
      STOP
      END
```

```
      SUBROUTINE GPLOT(X,Y,G,N,CH,NG)
      DIMENSION X(N),Y(N),F(1262),G(N),CH(NG),XX(2),MX(1),MY(1),CHAR(2)
      EQUIVALENCE (J,FJ)
      COMMON/GETCMD/NCHW,NCHA,NCHD,NCFN,IA,IC,IFMT(5),CFMT(5),NCHL,XOUT
      NCHW=4
      IA=6
      NCHL=132
      XMAX=X(1)
      XMIN=X(1)
      YMAX=Y(1)
      YMIN=Y(1)
      DO 10 I=2,N
      XMAX=AMAX1(X(I),XMAX)
      XMIN=AMIN1(X(I),XMIN)
      YMAX=AMAX1(Y(I),YMAX)
   10 YMIN=AMIN1(Y(I),YMIN)
      CALL PLTSTG(F,1262,IA,XMIN,XMAX,75,X,0,YMIN,YMAX,45,Y,0)
      MX(1)=1
      MY(1)=2
      CHAR(2)=CH(NG+1)
      DO 20 I=1,N
      J=G(I)
      IF(J.LE.0.OR.J.GT.NG) FJ=G(I)
      IF(J.LE.0.OR.J.GT.NG) GO TO 20
      CHAR(1)=CH(J)
      XX(1)=X(I)
      XX(2)=Y(I)
      CALL PLTAFL(F,XX,MX,MY,1,CHAR)
   20 CONTINUE
      CALL PLTPRT(1,F,0,F)
      RETURN
      END
```

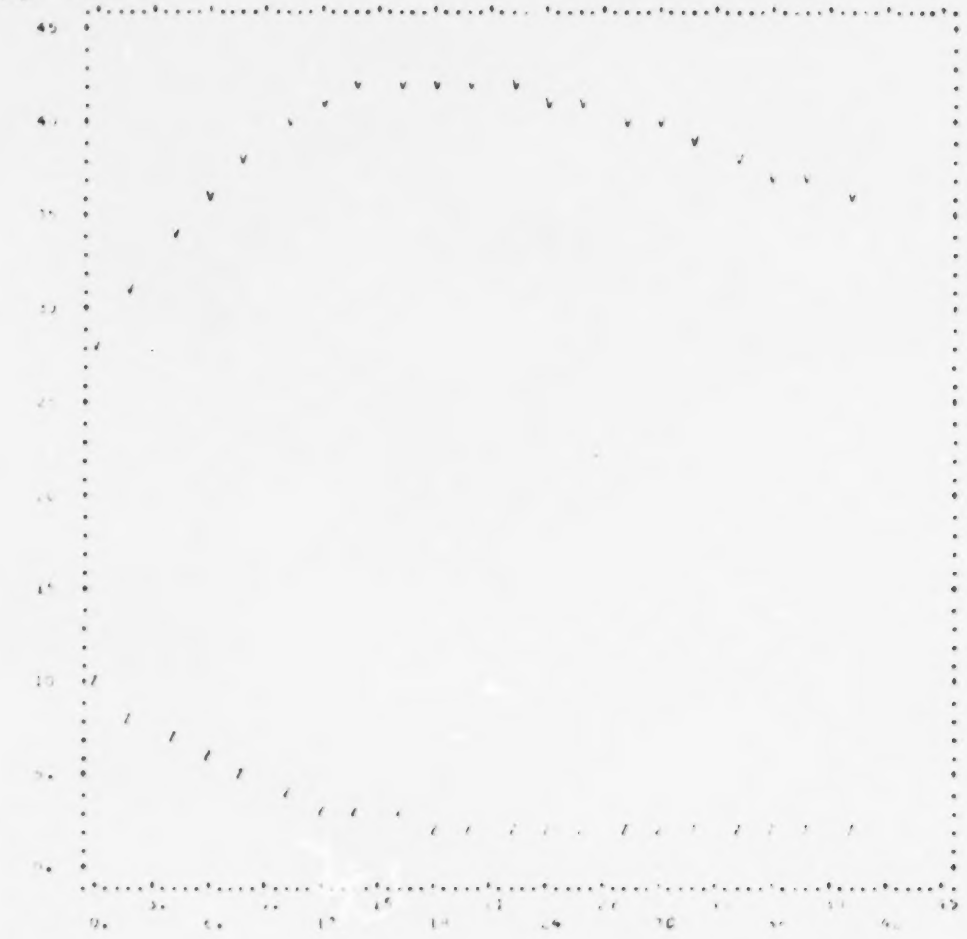
MODEL 1 A, B, C, D= 0.0 10.00000 0.00174 2.00040

I	TIME	V	Z
1	-0.0	19.37502	0.0
2	2.00000	17.50355	0.74333
3	4.00000	15.89048	1.55460
4	6.00000	14.46230	2.39002
5	8.00000	13.20042	3.24256
6	10.00000	12.15372	4.10819
7	12.00000	11.25141	4.98442
8	14.00000	10.43222	5.86955
9	16.00000	9.70174	6.76238
10	18.00000	9.05239	7.66195
11	20.00000	8.47264	8.56708
12	22.00000	7.95308	9.47800
13	24.00000	7.48304	10.39500
14	26.00000	7.06355	11.31569
15	28.00000	6.68110	12.24101
16	30.00000	6.33339	13.17013
17	32.00000	6.01620	14.10301
18	34.00000	5.72620	15.03943
19	36.00000	5.46000	15.97910
20	38.00000	5.21928	16.92200
21	40.00000	4.99550	17.86797

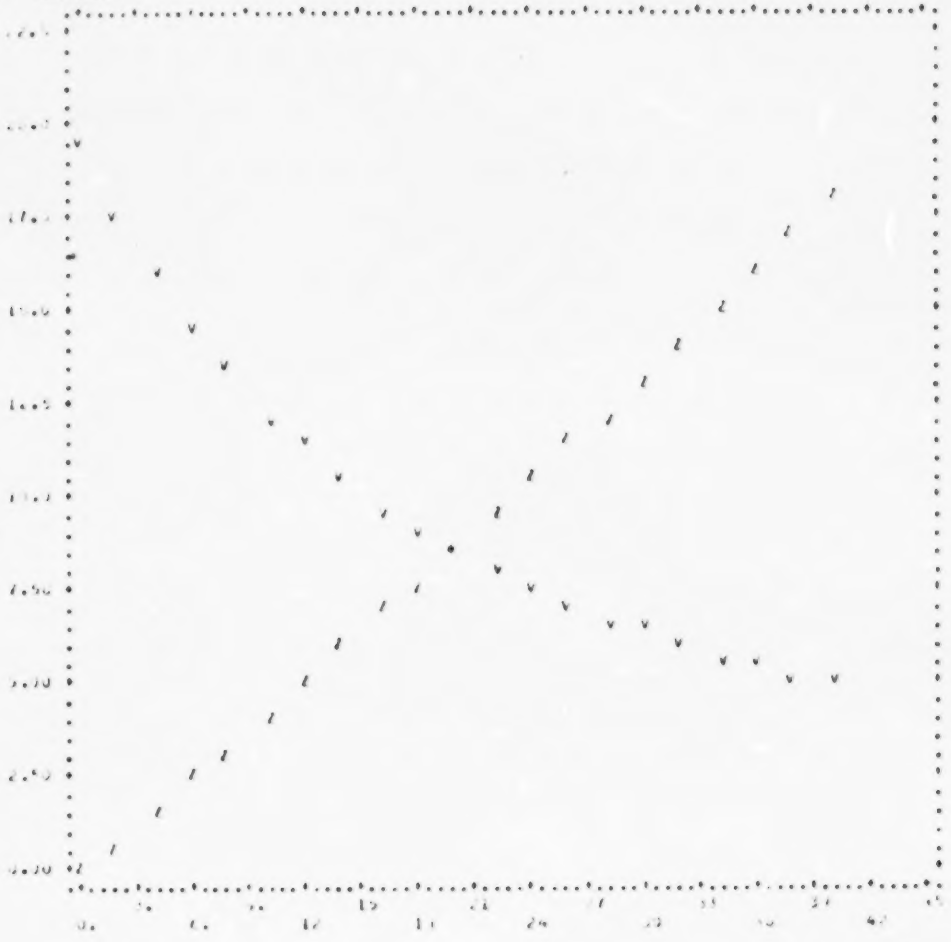
MODEL 2 A, B, C, D= 1.00000 0.10000 0.00020 2.00000

I	TIME	V	Z
1	-0.0	27.75547	10.00000
2	2.00000	31.11500	8.25731
3	4.00000	34.02652	6.66319
4	6.00000	35.45024	5.72814
5	8.00000	36.38330	4.81329
6	10.00000	34.85205	4.07879
7	12.00000	40.50156	3.49194
8	14.00000	41.58597	3.02597
9	16.00000	41.55340	2.65896
10	18.00000	42.00000	2.37295
11	20.00000	41.96121	2.15335
12	22.00000	41.65055	1.98803
13	24.00000	41.20172	1.86716
14	26.00000	40.76327	1.76275
15	28.00000	40.29367	1.72810
16	30.00000	39.50220	1.69767
17	32.00000	38.85558	1.68762
18	34.00000	38.19852	1.69375
19	36.00000	37.48705	1.71324
20	38.00000	36.76951	1.74371
21	40.00000	36.05350	1.78316

Z,V VERSUS T



Z ,V VERSUS T



CHAPTER 6

CONCLUDING REMARKS

During the next twelve months, the marginal and combined effects of the basic pollutants on morbidity for the City of Los Angeles and the mile-high city of Denver, Colorado, will be assessed using the methods developed in this report. The various ways for modeling for the competing risk situation will be explored. Both parametric and nonparametric methods will be considered. We will show how prior knowledge of empirical distributions of the concentrations of the various pollutants can be used to sharpen the statistical procedures of estimation and testing for goodness-of-fit as a follow-on effort, including applications of paramount importance to the Environmental Protection Agency of the United States.

CHAPTER 7

Bibliography

Armitage, P. [1959]
"The Comparison of Survival Curves," Journal of the Royal Statistical Society, A, Part III, pp. 12-294.

Bain, L.J.; Eastman, J.; and Engelhardt, M. [February 1973]
"A Study of Life-Testing Models and Statistical Analysis for the Logistic Distribution," Aerospace Research Laboratories Report ARL 73-0009, Wright-Patterson Air Force Base, Ohio.

Berg, J.W. and Robbins, G.F. [1967]
"The Failure of a Model to Predict Cancer Survival," Journal of Chronic Diseases, Vol. 20, pp. 809.

Berkson, J. and Gage, R.P. [1952]
"Survival Curve for Cancer Patients Following Treatment," Journal of the American Statistical Association, Vol. 47, pp. 501-515.

Boag, J.W. [1949]
"Maximum Likelihood Estimates of the Proportion of Patients Cured by Cancer Therapy," Journal of the Royal Statistical Society, B. Vol. 11, pp. 15-53.

Broadbent, S. [1958]
"Simple Mortality Rates," Applied Statistics, Vol. 7, pp. 86-95.

Buckland, W.R. [1964]
Statistical Assessment of the Life Characteristic. Griffin's Statistical Monograph, No. 13. New York: Hafner.

Chanda, K.C. [1954]
"A Note on the Consistency and Maxima of the Roots of Likelihood Equations," Biometrika, Vol. 41, pp. 96-61.

Chiang, C.L. [1961]
"A Stochastic Study of the Life Table and Its Applications," Biometrics, Vol. 17, pp. 57-78.

David, M.A. [1970]
"On Chiang's Proportionality Assumption in the Theory of Competing Risks, Note 282," Biometrics, Vol. 26, pp. 336-339.

Davis, D.J. [1952]
"An Analysis of Some Failure Data," Journal of the American Statistical Association, Vol. 47, pp. 113-149.

Dixon, W.J.; Longmire, W.P.; and Holden, W.D. [January 1971]
"Use of Triethylenethiophosphoramide (thioTEPA) as an Adjuvant to the Surgical Treatment of Gastric and Colorectal Carcinoma: Ten-Year Follow-Up," Annals of Surgery, Vol. 173, No. 1, pp. 26-39.

Ederer, F.; Axtell, L.M.; and Cutler, S.J. [Undated]
"The Relative Survival Rate: A Statistical Methodology," in Cancer: End Results and Mortality Trends, National Cancer Institute Monograph No. 6, U.S. Department of Health, Education and Welfare, National Institute of Health.

Epstein, B. and Sobel, M. [1953]
"Life Testing," Journal of the American Statistical Association, Vol. 48, pp. 486-502.

Fix, E. and Neyman, J. [1951]
"A Simple Stochastic Model of Recovery, Relapse, Death and Loss of Patients," Human Biology, Vol. 23, pp. 205-241.

Garg, M.L.; Rao, B.R.; and Redmond, C.K. [1970]
"Maximum-Likelihood Estimation of the Parameters of the Gompertz Survival Function," Journal of the Royal Statistical Society, Applied Statistics, C, Vol. 19, pp. 152-159.

Gehan, E.A. [1969]
"Estimating Survival Functions From the Life Table," Journal of Chronic Diseases, Vol. 21, pp. 629-644.

Ghare, P.M. and Kim, Y.H. [1970]
An Approximating Function for the Hazard Rate. Pittsburgh, Penn.: 1970 Technical Conference Transactions, American Society of Quality Control, pp. 367-373.

Gompertz, E. [1825]
"On the Nature of the Function Expressive Law of Human Mortality and on the New Mode of Determining the Value of Life Contingencies," Phil. Trans. Royal Soc., A, Vol. 115, pp. 513.

Grenander, U. [1956]
"On the Theory of Mortality Measurement," Skandinavisk Aktuarietidskrift, Vol. 39, pp. 1-55.

Halperin, M. [1952]
"Maximum Likelihood Estimation in Truncated Samples," Annals of Mathematical Statistics, Vol. 23, pp. 226-238.

Harris, T.E.; Meier, P.; and Tukey, J.W. [December 1950]
"Timing of the Distribution of Events Between Observations," Human Biology, Vol. 22, No. 4.

Kale, B.K. [1962]
"On the Solution of Likelihood Equations by Iterations Processes. The Multiparametric Case," Biometrika, Vol. 49, pp. 479-486.

Kimball, A.W. [1960]
"Estimation of Mortality Intensities in Animal Experiments," Biometrics, Vol. 16, pp. 505-521.

Kodlin, D. [1967]
"A New Response Time Distribution," Biometrics, Vol. 23, pp. 227-330.

Mantel, N. [1966]
"Evaluation of Survival Data and Two New Rank Order Statistics Arising in its Consideration," Cancer Chemotherapy Reports, Vol. 50, pp. 163-170.

Moeschberger, M.L. and David, H.A. [1971]
"Life Tests Under Competing Causes of Failure and the Theory of Competing Risks," Biometrics, Vol. 27, pp. 909-933.

Murthy, V.K. [1965a]
"Estimation of Probability Density," Annals of Mathematical Statistics, Vol. 36, pp. 1027-1031.

Murthy, V.K. [1965b]
"Estimation of Jumps, Reliability and Hazard Rate," Annals of Mathematical Statistics, Vol. 36, pp. 1032-1040.

Murthy, V.K. [1968]
"Study of Nonparametric Techniques for Estimating Reliability and Other Quality Parameters," NASA Contract Final Report, McDonnell-Douglas Corporation paper.

Murthy, V.K. and Swartz, G.B. [March 1971]
"Contributions to Reliability," Aerospace Research Laboratories Report ARL 71-0060, Wright-Patterson Air Force Base, Dayton, Ohio.

Sampford, M.R. [1952]
"The Estimation of Response-Time Distributions. I. Fundamental Concepts and General Methods; II. Multi-Stimulus Distributions," Biometrics, Vol. 8, pp. 13-32 and 307-369.

Swartz, G.B. [June 1973]
"The Mean Residual Lifetime Function," IEEE Transactions on Reliability, pp. 108-109.

Watson, G.S. and Leadbetter, M.R. [1964]
"Hazard Analysis, I," Biometrika, Vol. 51, pp. 175-184.

Zelen, M. [1966]
"Application of Exponential Models to Problems in Cancer Research," Journal of the Royal Statistical Society, A, Vol. 129, pp. 368-398.